Break Through Science on a Petaflop XT5

April, 2009

John M Levesque

CTO Office

Applications

Cray Supercomputing Center of Excellence

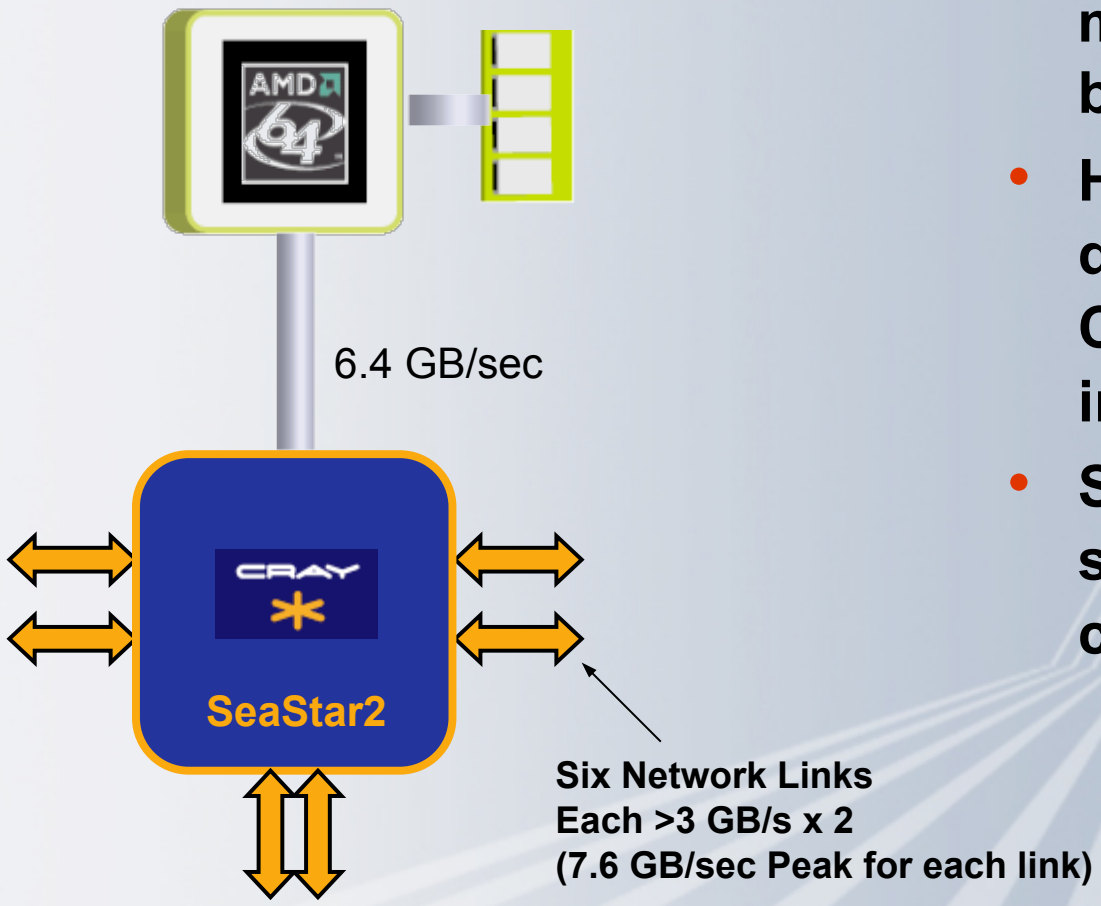
CRAY
THE SUPERCOMPUTER COMPANY

Recipe for a good MPP

1. Select Best Microprocessor
2. Surround it with a balanced or “bandwidth rich” environment
3. “Scale” the System
   - Eliminate Operating System Interference (OS Jitter)
   - Design in Reliability and Resiliency
   - Provide Scaleable System Management
   - Provide Scalable I/O
   - Provide Scalable Programming and Performance Tools
   - System Service Life (provide an upgrade path)

AMD Opteron: Why we selected it

RAY XT4 PE

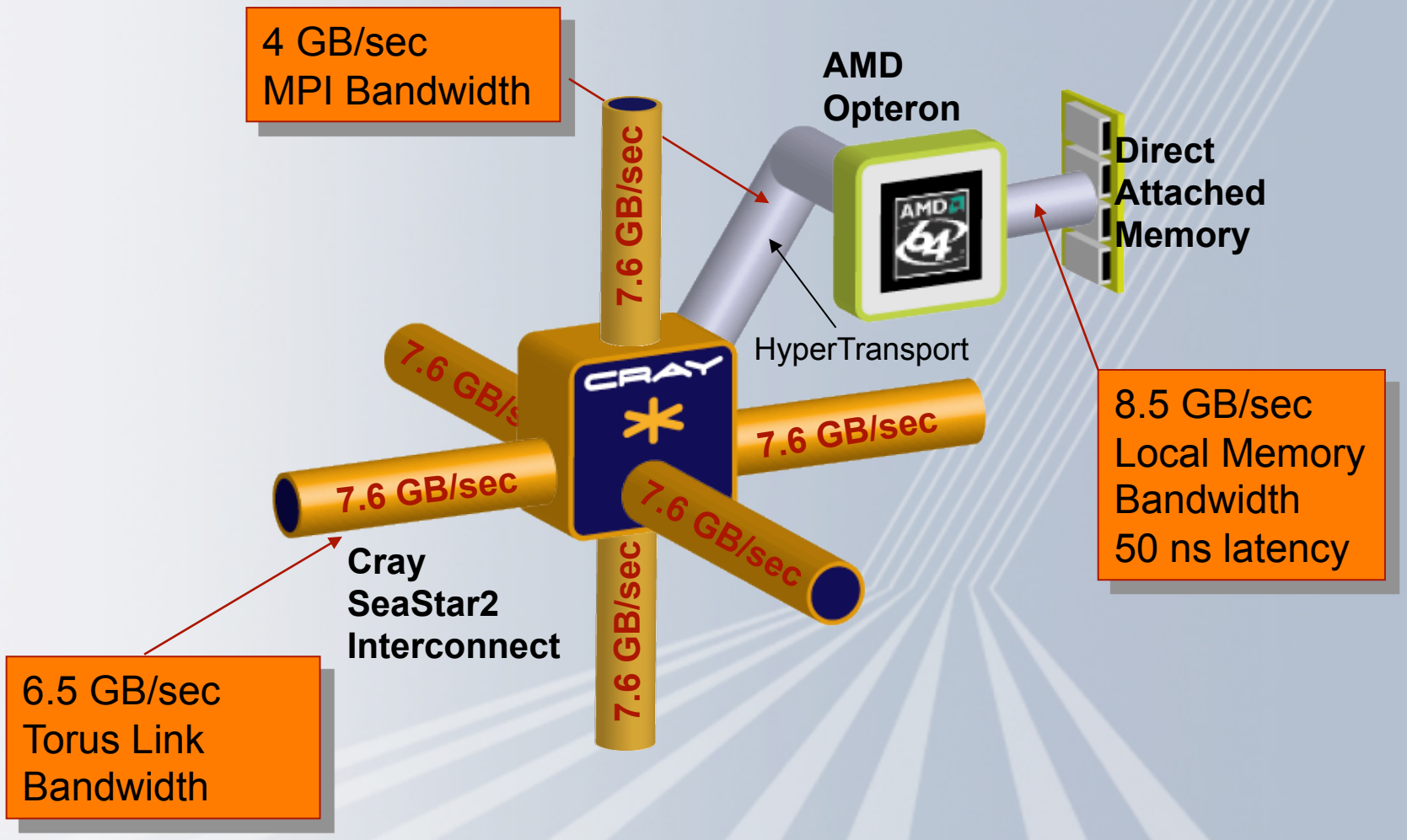


- Direct attached local memory for leading bandwidth and latency
- HyperTransport can be directly attached to Cray SeaStar2 interconnect
- Simple two-chip design saves power and complexity

Recipe for a good MPP

1. Select Best Microprocessor
2. Surround it with a balanced or “bandwidth rich” environment
3. “Scale” the System
   - Eliminate Operating System Interference (OS Jitter)
   - Design in Reliability and Resiliency
   - Provide Scalable System Management
   - Provide Scalable I/O
   - Provide Scalable Programming and Performance Tools
   - System Service Life (provide an upgrade path)

The Cray XT4 Processing Element: Providing a bandwidth-rich environment

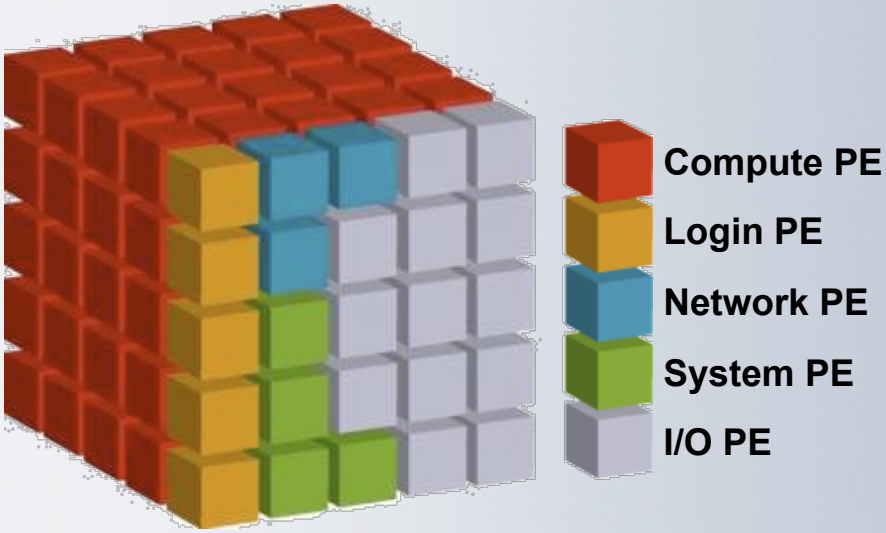


Recipe for a good MPP

1. Select Best Microprocessor
2. Surround it with a balanced or “bandwidth rich” environment
3. “Scale” the System
   - Eliminate Operating System Interference (OS Jitter)
   - Design in Reliability and Resiliency
   - Provide Scalable System Management
   - Provide Scalable I/O
   - Provide Scalable Programming and Performance Tools
   - System Service Life (provide an upgrade path)

Scalable Software Architecture: UNICOS/lc

"*Primum non nocere*"



Service Partition

Specialized Linux nodes

- Microkernel on Compute PEs, full featured Linux on Service PEs.
- Service PEs specialize by function
- Software Architecture eliminates OS "Jitter"
- Software Architecture enables reproducible run times
- Large machines boot in under 30 minutes, including filesystem

This is the real reason the XT4 will scale to a Petaflop

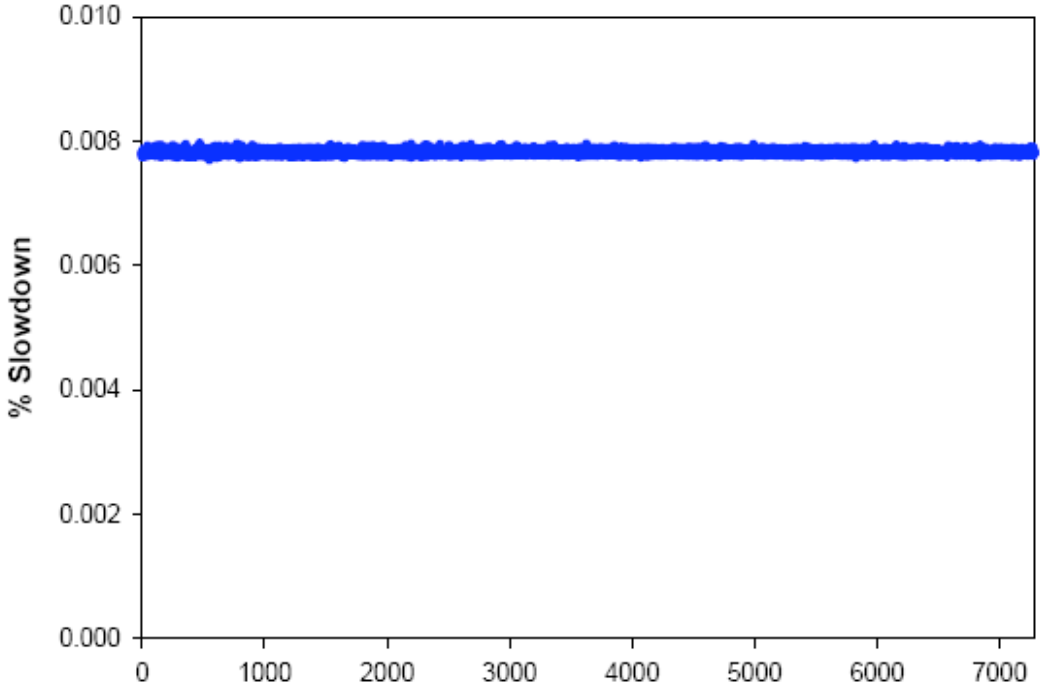
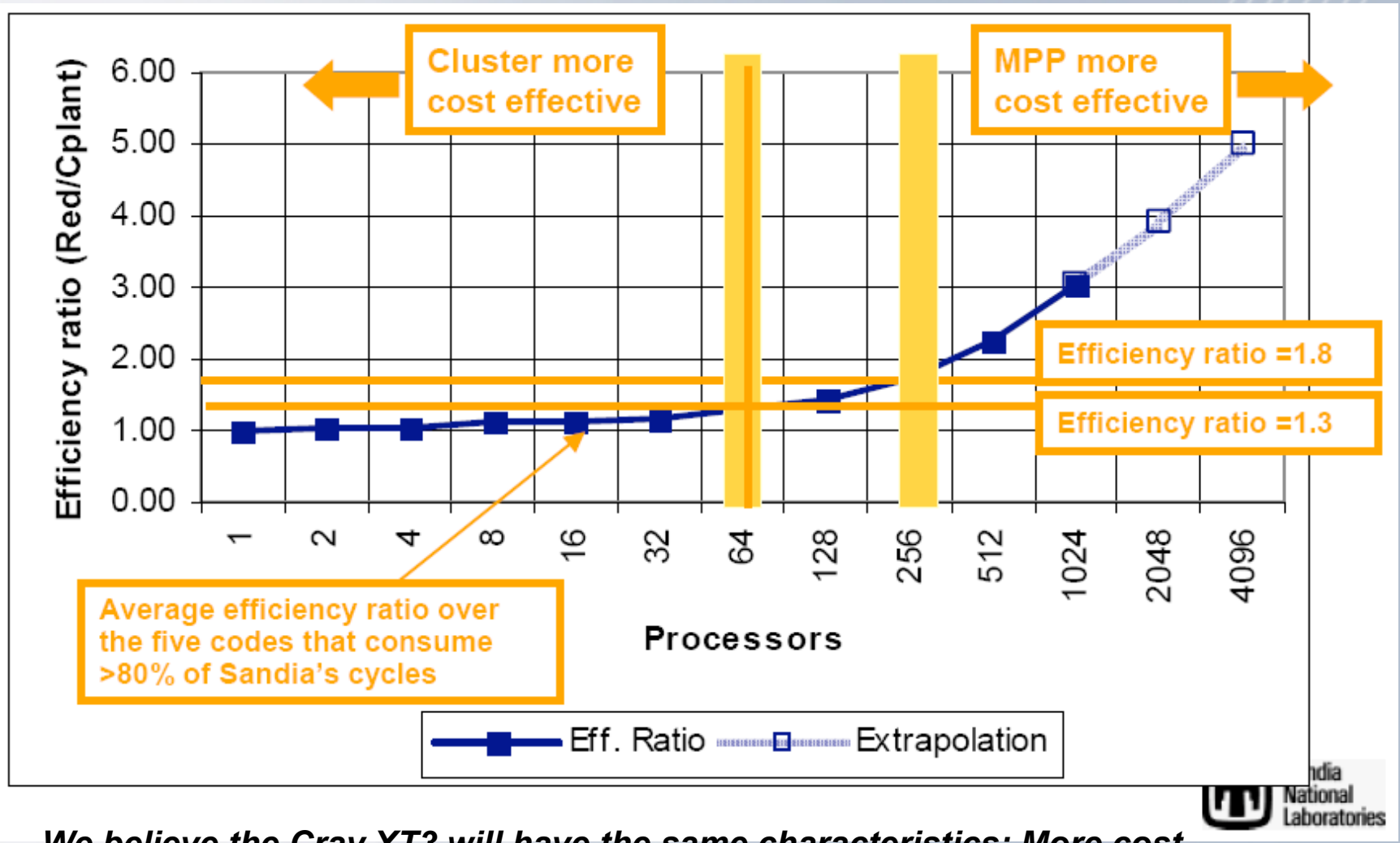


Figure 3: Slowdown caused by computational "noise"

Download P-SNAP from the web and try it on your system

Relating Scalability and Cost Effectiveness of Red Storm Architecture

Source: Sandia National Labs



We believe the Cray XT3 will have the same characteristics; More cost effective than clusters somewhere between 64 and 256 MPI tasks

AMD Core Optimizations

- Cache Optimizations
  - Cache Associativity and Cache Thrashing
  - Cache Alignment
  - Cache Blocking
  - OpenMP
- Prefetching
  - Types of Prefetching
  - When to use
- Programmatic Examples
  - Striding
  - Function Calls
  - Importance of Vectorization
  - Matrix Multiplies
  - Loop Optimizations
- Things to remember

Let's Review: Dual Core v. Quad Core

Dual Core

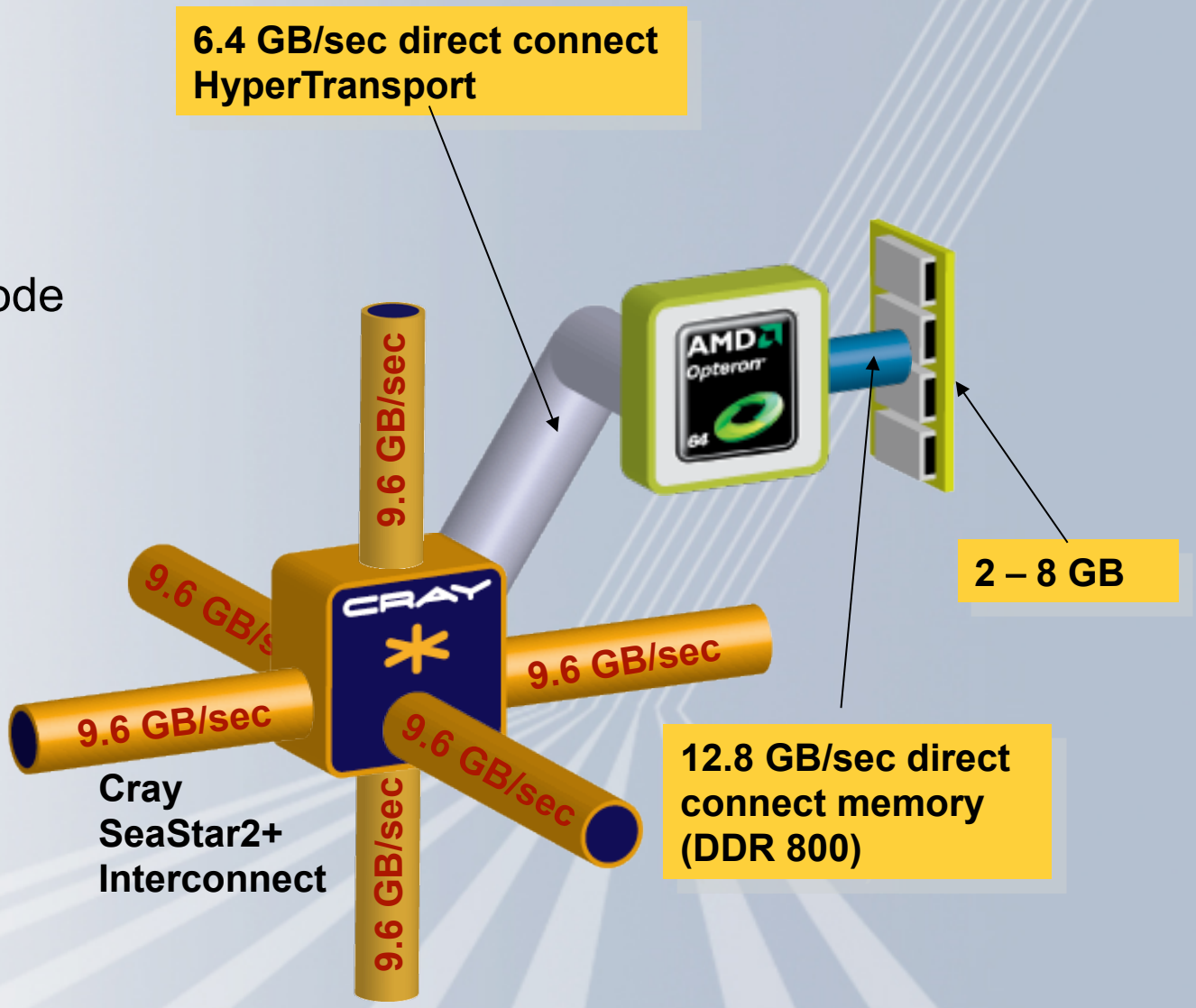
- Core
  - 2.6Ghz clock frequency
  - SSE SIMD FPU (2flops/cycle = 5.2GF peak)
- Cache Hierarchy
  - L1 Dcache/Icache: 64k/core
  - L2 D/I cache: 1M/core
  - SW Prefetch and loads to L1
  - Evictions and HW prefetch to L2
- Memory
  - Dual Channel DDR2
  - 10GB/s peak @ 667MHz
  - 8GB/s nominal STREAMs

Quad Core

- Core
  - 2.1Ghz clock frequency
  - SSE SIMD FPU (4flops/cycle = 8.4GF peak)
- Cache Hierarchy
  - L1 Dcache/Icache: 64k/core
  - L2 D/I cache: 512 KB/core
  - L3 Shared cache 2MB/Socket
  - SW Prefetch and loads to L1,L2,L3
  - Evictions and HW prefetch to L1,L2,L3
- Memory
  - Dual Channel DDR2
  - 12GB/s peak @ 800MHz
  - 10GB/s nominal STREAMs

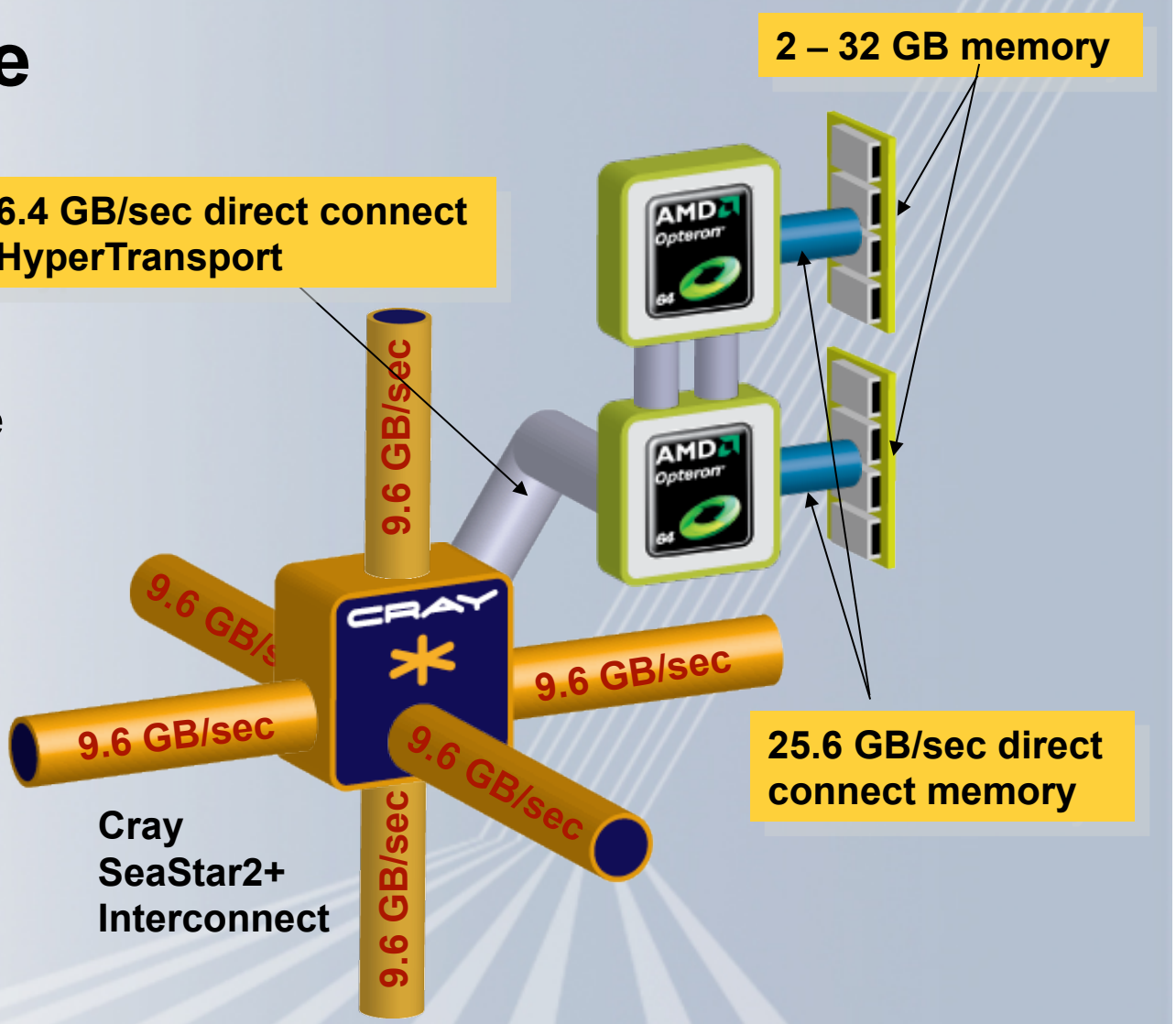
Cray XT4 Node

- 4-way SMP
- >35 Gflops per node
- Up to 8 GB per node
- OpenMP Support within socket

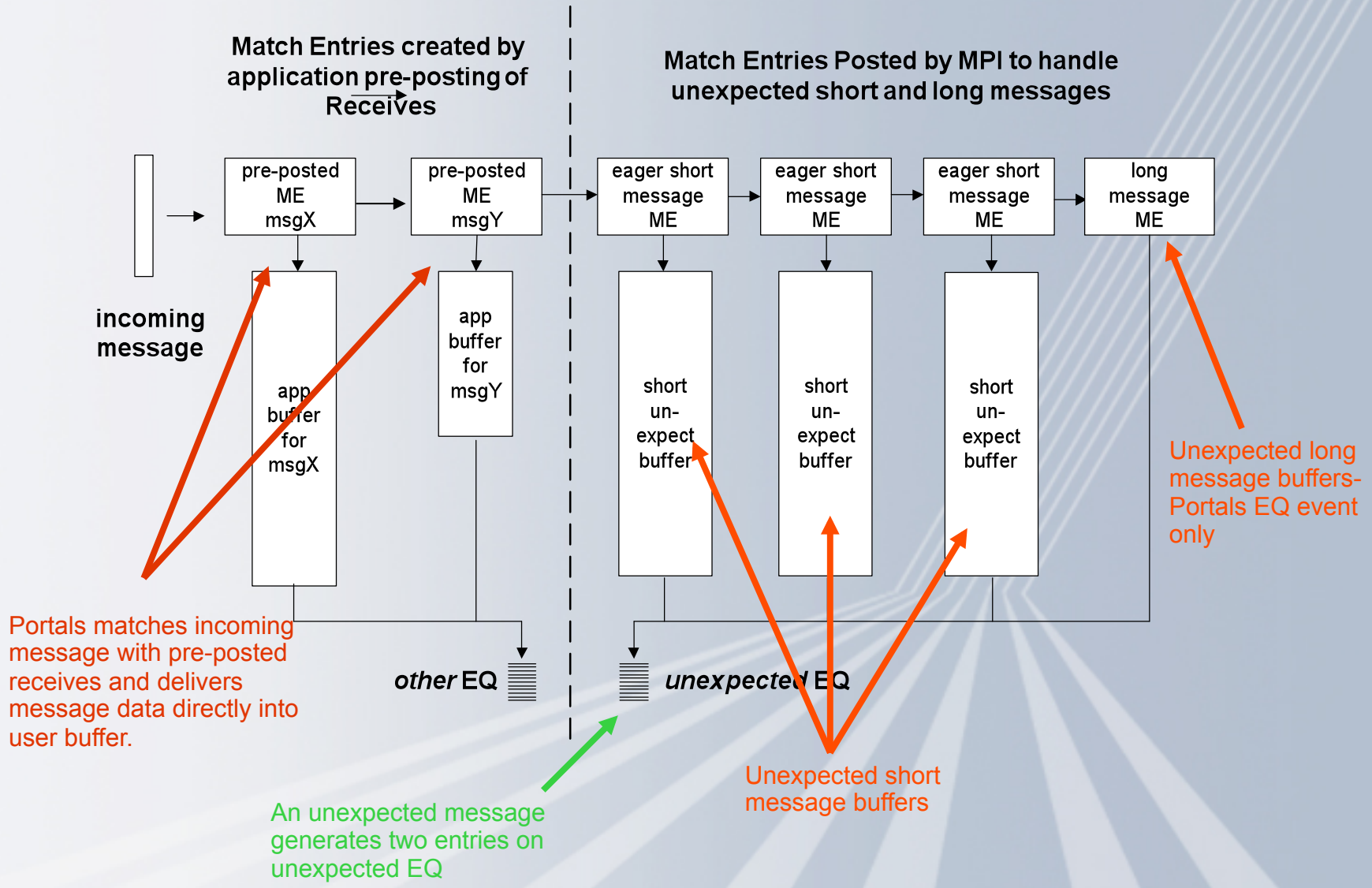


Cray XT5 Node

- 8-way SMP
- >70 Gflops per node
- Up to 32 GB of shared memory per node
- OpenMP Support



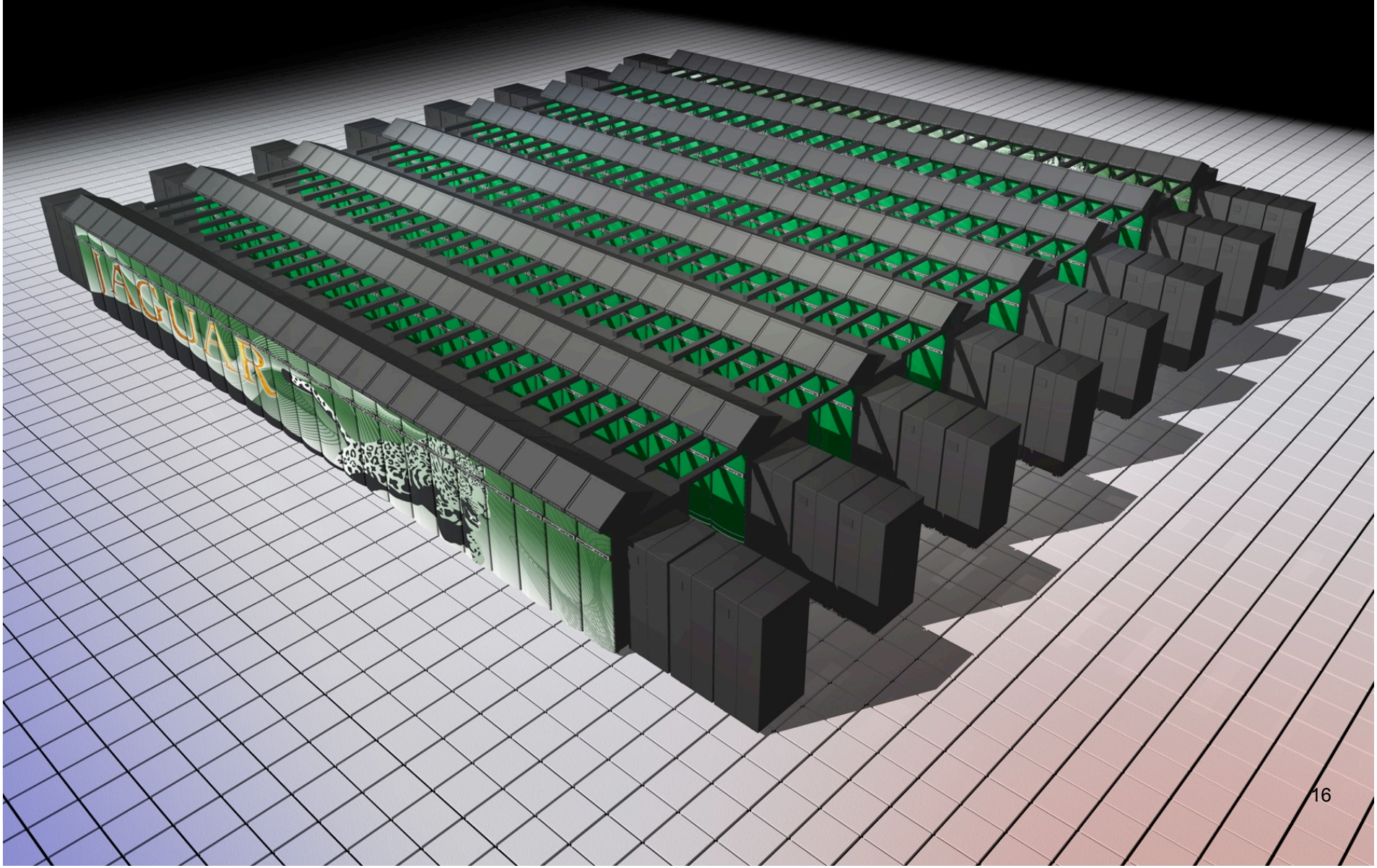
XT MPI – Receive Side



Short History of Jaguar Petaflop

- System completely installed on the floor at ORNL the first week of October. 200 new liquid cooled cabinets
  - Amazing ramp up of Cray manufacturing to make this happen
- Used HPL to shake down the machine
  - Get a long enough run on a brand new system to place high in Top 500
- Ran a few applications between HPL runs
  - Amazed at scalability
- November 7th we started running applications
  - Eight applications, from 5 different science areas set World records on performance
- November 10th SPECfem3d ported and ran on 149784 cores
  - Actually hit a code error, because a problem of this size and resolution had never been tried before. This was fixed quickly and the run succeeded.
- November 12th Lin-Wang Wang code from UCB ported and run on 149144 cores
- Made numerous HPCC runs this last week

ORNL Petaflops System



Jaguar XT5 will be integrated with the XT4 after acceptance – 1.6PF's and 284 cabinets!

Jaguar	Total	XT5	XT4
Peak Performance	**1,645**	**1,382**	**263**
AMD Opteron Cores	181,504	150,176	31,328
System Memory (TB)	362	300	62
Disk Bandwidth (GB/s)	284	240	44
Disk Space (TB)	10,750	10,000	750
Interconnect Bandwidth (TB/s)	532	374	157

Two Months and potential break-through Science for 8 research groups

All of these runs set new World Records in Performance

Science Area	Code	Contact	Cores	% of Peak	Total Perf	Notes	Scaling
Materials	DCA++	Schulthess	150144	97%	1.35 PF	2008 Gordon Bell Finalist	Weak
Materials	LSMS/WL	ORNL	149580	76.40%	1.05 PF	64 bit	Weak
Seismology	SPECFEM3D	UCSD	149784	12.60%	164 TF	2008 Gordon Bell Finalist	Weak
Weather	WRF	Michalakes	150000	3.60%	50 TF	2007 Gordon Bell Finalist	Strong
Climate	POP	Jones	21000		20 sim yrs/ CPU dau	Size of Data	Strong
Combustion	S3D	Chen	144000	6.00%	83 TF		Weak
Fusion	GTC	PPPL	102000		20 billion particles pushed	Code Limit	Weak
Materials	LS3DF	Lin-Wang Wang	147456	32%	425 TF	2008 Gordon Bell Finalist	Weak

* Mixed Precision – 626 TF at 128K cores in 64 bit only

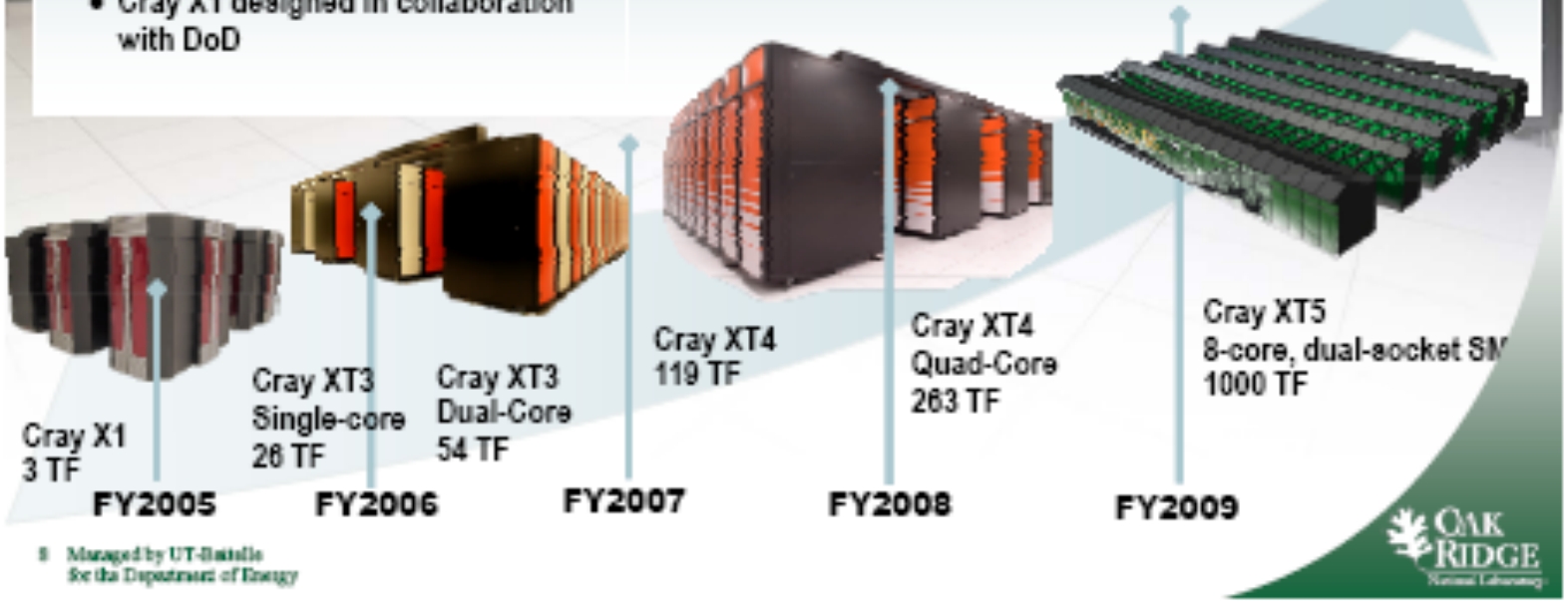
It Does Help working with some of the best researchers in the Country – the DoE Office of Science Incite scientists

Lets Look at what
Oak Ridge National Laboratory
has put together

We have increased system performance by more than 300 times since 2004

Hardware scaled from single-core through dual-core to quad-core and dual-socket SMP nodes	Scaling applications and system software is the biggest challenge

• NNSA and DoD have funded much of the basic system architecture research
  • Cray XT based on Sandia Red Storm
  • IBM BG designed with Livermore
  • Cray X1 designed in collaboration with DoD
	• SciDAC program is funding scalable application work that has advanced many science apps
	• DOE-SC and NSF have funded much of the library and applied math as well as tools
	• Computational Liaisons key to using deplo[illegible]ems



8 Managed by UT-Battelle for the Department of Energy

Why are DoE Office of Science Researchers so well prepared for Petaflops at ORNL?

- For the past fours years Cray has worked with DoE Office of Science developers and the Computer Science Group at ORNL to prepare their applications for larger and larger processor counts
- They also have some of the best application developers in the industry

POC
Thomas
Schulthess, ORNL

Pioneering Application: DCA++

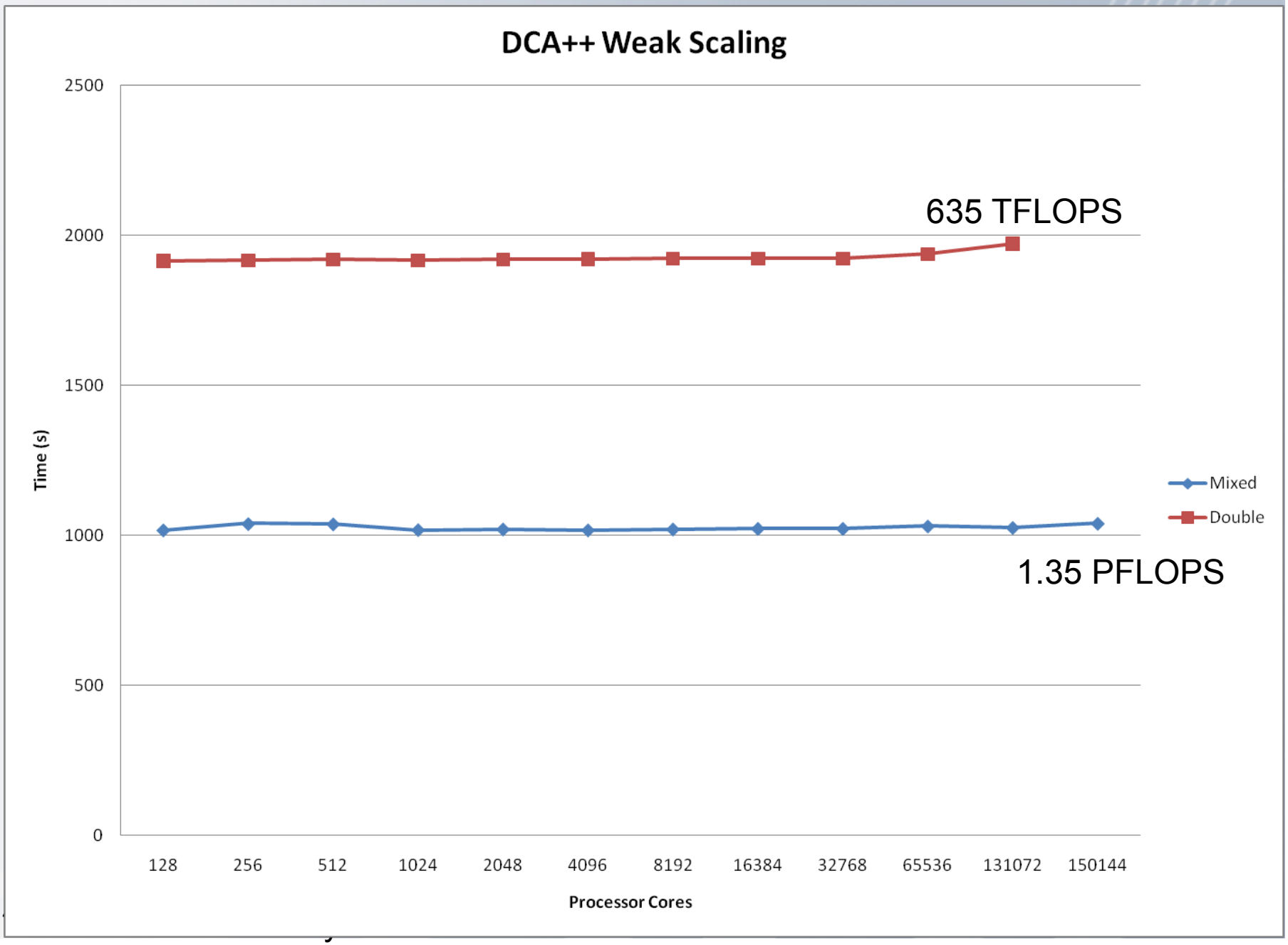
Science Goals and Impact

Science Goals

- Study high temperature superconductivity (HTC) via simulations of inhomogeneous Hubbard models
  - Believed to describe the HTC cuprates
- Recent simulations have shown that the 2D homogeneous Hubbard model does have a superconducting state and pairing mechanism is now understood
  - The responsible pairing interaction arises from anti-ferromagnetic spin fluctuations
- Must address the effect of charge & spin inhomogeneities on the superconducting state in the Hubbard model
  - Their effect on the critical temperature Tc and their role in the pairing mechanism
- Studies of both random and periodic inhomogeneities will be carried out

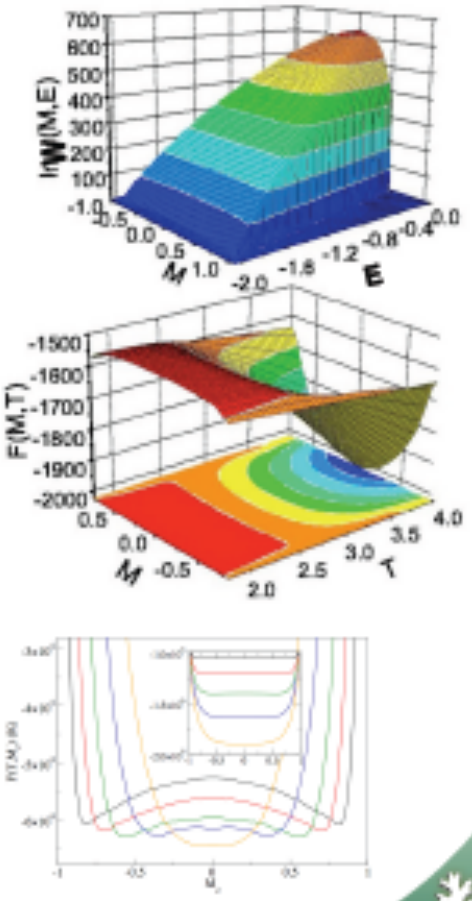
Science Impact

- Recent experiments have shown that nanoscale charge and spin inhomogeneities emerge in a number of cuprates
- Based on these findings, it was proposed in the literature that inhomogeneities play a major role in HTC
- Results will be used to study the role of inhomogeneities in the pairing mechanism of the 2D Hubbard model and address questions such as
  - Do inhomogeneities act to increase or decrease the critical temperature Tc?
  - Do they enhance, suppress or even modify the pairing mechanism?
  - Is there an optimal inhomogeneity that maximizes Tc?
- Use the knowledge gained to artificially structure cuprate based materials with higher transition temperatures


DCA++ Weak Scaling
2500
2000
1500
1000
500
0
Time (s)
635 TFLOPS
1.35 PFLOPS
Mixed
Double
128
256
512
1024
2048
4096
8192
16384
32768
65536
131072
150144
Processor Cores

Wang-Landau/LSMS hybrid code

- The Wang-Landau (WL) algorithm with Frontier sampling allows us to compute Free energies in atomistic simulations.
- For example, we are able to compute the temperature dependent free energy barrier for magnetization reversal of magnetic nanoparticles.
  - Figures show simulations of the temperature dependent free energy as a function of magnetization for a simple model of FePt nanoparticles.
- In the WL/LSMS code allows us to use ab-initio electronic structure simulations as the underlying energy function
  - Simulations a now quantitative and sensitive to the chemical composition of the nanoparticles. This will allow us to study the thermal behavior of the magnetization in individual nanoparticles - a task that is almost impossible with today's experimental techniques.



LSMS-WL

- Achieved 76.4% of peak on 149,508 cores =1.05 PF

WRF Nature Run

John Michalakes
Josh Hacker
Richard Loft

National Center for Atmospheric Research (NCAR), Boulder, CO.

Michael O.McCracken
Allan Snavely
Nicholas J. Wright

Performance Modeling and Characterization Lab San Diego Supercomputer Center, La Jolla, CA.

Tom Spelce
Brent Gorda

Lawrence Livermore National Laboratory, Livermore, CA.

Robert Walkup

IBM Thomas J. Watson Research Center, Yorktown Heights, NY.

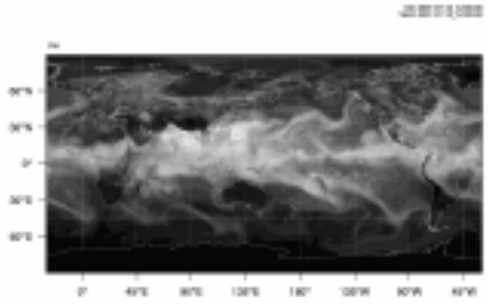
Outline

- Statement of problem and goals
- WRF overview and petascale issues
- Nature run methodology
- Results and conclusion

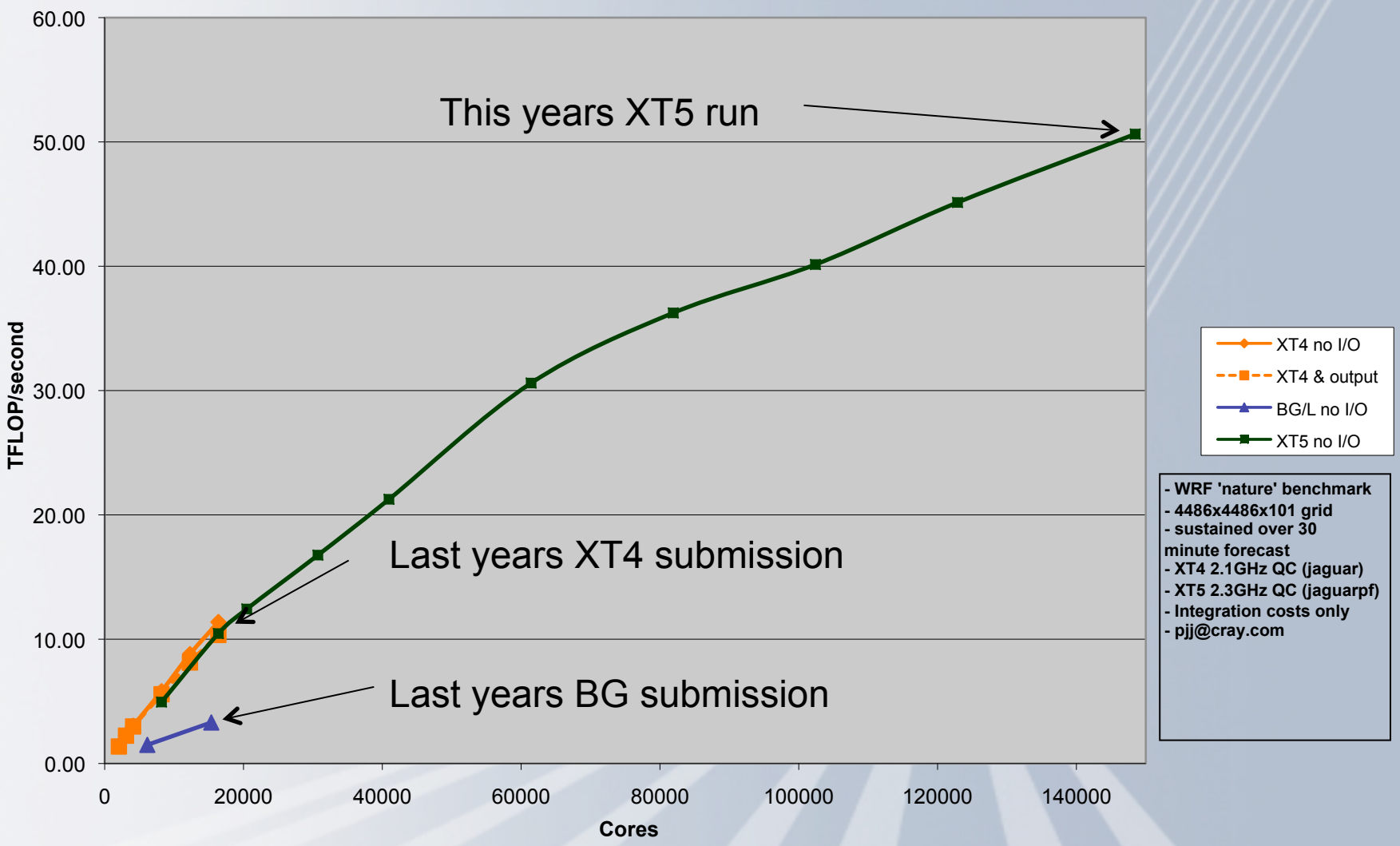
WRF Overview

- Large collaborative effort to develop next-generation community model with direct path to operations
  - Limited area, high-resolution
  - Structured (Cartesian) with mesh-refinement (nesting)
  - High-order explicit dynamics
  - Software designed for HPC
  - 3000+ registered users
- Applications
  - Numerical Weather Prediction
  - Atmospheric Research
  - Coupled modeling systems
  - Air quality research/prediction
  - High resolution regional climate
  - Global high-resolution WRF

http://www.wrf-model.org



5 day global WRF forecast at 20km horizontal resolution, running at 4x real time 128 processors of IBM Power5+ (bluefire.ucar.edu)


WRF 'nature' benchmark on Cray XT
60.00
50.00
40.00
30.00
20.00
10.00
0.00
TFLOP/second
0
20000
40000
60000
80000
100000
120000
140000
Cores
This years XT5 run
Last years XT4 submission
Last years BG submission
XT4 no I/O
XT4 & output
BG/L no I/O
XT5 no I/O
- WRF 'nature' benchmark
- 4486x4486x101 grid
- sustained over 30 minute forecast
- XT4 2.1GHz QC (jaguar)
- XT5 2.3GHz QC (jaguarpf)
- Integration costs only
- pjj@cray.com

POC
Jackie Chen, SNL

Pioneering Application: S3D

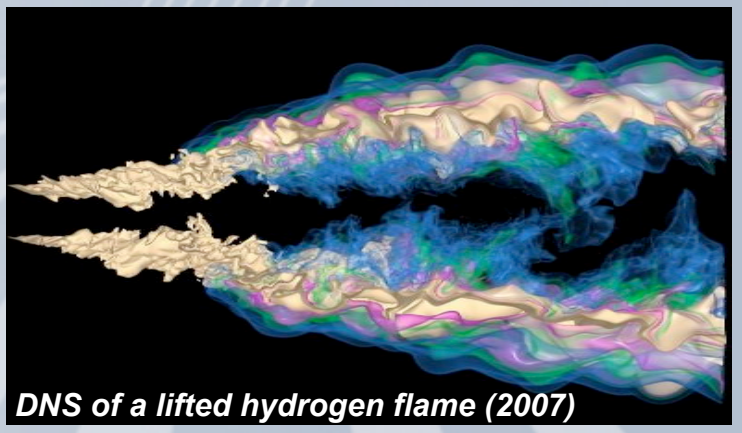
Science Goals and Impact

Science Goals

- Turbulent lifted flames occur in diesel engines and gas turbines
  - Fuel is injected into a hot gas environment and flame is stabilized through the recirculation of hot air and combustion products
- What are the mechanisms that stabilize the flame base?
  - Explore the role of auto-ignition, flame propagation, and large eddies
- Analyze a lifted turbulent slot jet flame with a heated coflow
  - Extend a recent H2/air lifted jet flame configuration in ambient coflow to more realistic chemistry (ethylene) and higher pressures representative of compression ignition engine operating regimes
- Detailed of proposed simulation
  - 15 um grid spacing, 2 mm nozzle jet height, 2.4 cm axial length, 3.2 cm transverse width, 0.6 cm spanwise
  - 200 m/s jet velocity (Re = 11,000)
  - Simulate 3 flow-through times (0.36 ms) for stationary statistics at lifted flame base

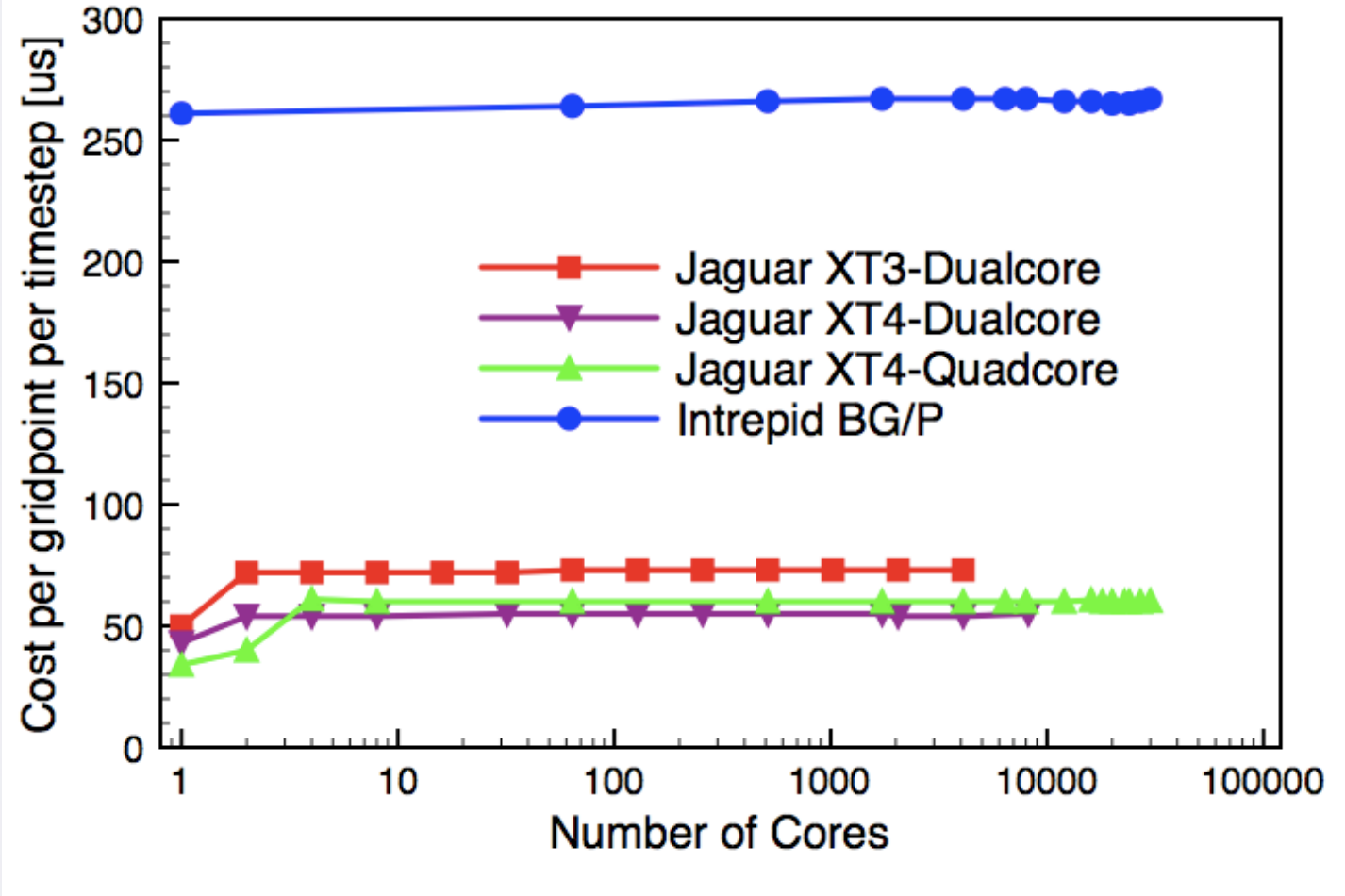
Science Impact

- Fundamental insight into lifted-flame stabilization mechanisms in auto-ignitive environments
- Provision of data for ignition and combustion model validation
- Acceleration of the evolution of a validated, predictive, multiscale, combustion modeling capability
- Optimize design and operation of evolving fuels in advanced engines for transportation applications.

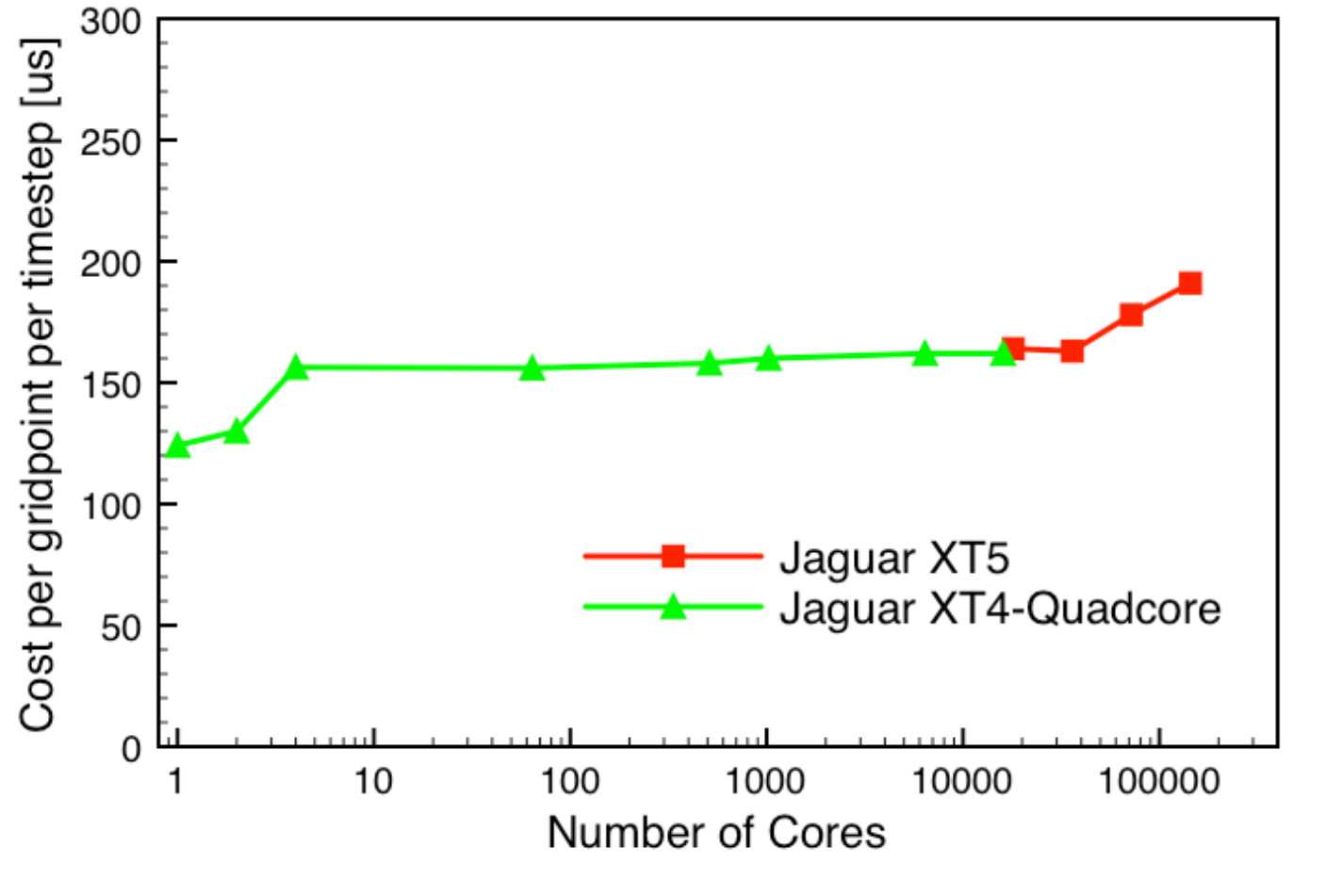


DNS of a lifted hydrogen flame (2007)

COH2 -


Cost per gridpoint per timestep [us]
300
250
200
150
100
50
0
1
10
100
1000
10000
100000
Number of Cores
Jaguar XT3-Dualcore
Jaguar XT4-Dualcore
Jaguar XT4-Quadcore
Intrepid BG/P

C2H4


Cost per gridpoint per timestep [us]
300
250
200
150
100
50
0
1
10
100
1000
10000
100000
Number of Cores
Jaguar XT5
Jaguar XT4-Quadcore

Pioneering Application: GTC

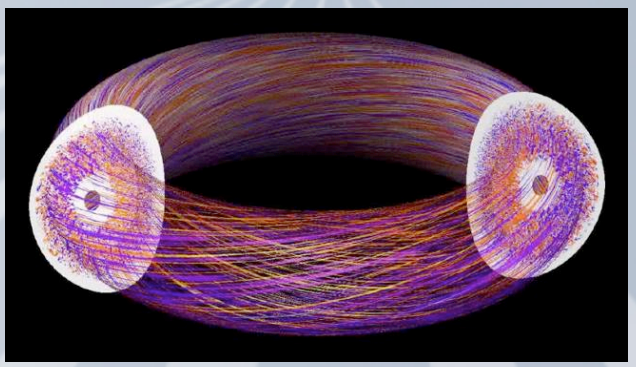
Physical Models and Algorithms

Physical Models

- GTC is a global code for turbulence transport simulations
  - Uses a shaped plasma in general geometry with electrostatic electron dynamics based on the δh scheme for nonadiabatic part of δf
- Based on the Particle-In-Cell method for solving the gyrokinetic Vlasov-Maxwell equations.
- GTC-C version of GTC uses a circular cross-section model geometry in the large-aspect ratio limit and can accommodate both kinetic ions & electrons
- GTC-S version of GTC can simulate more realistic plasmas where shaping effects are important
  - Global general geometry interfaced with realistic fusion plasma experimental profiles through the TRANSP fusion data tool

Numerical Algorithms

- Gyrokinetic Vlasov equation is solved with standard PIC method
  - Scatter-and-add operation is used for charge and current deposition on the grid
  - Gather operation is used to calculate the fields associated with each particle
- Gyrokinetic Poisson's equation and the associated continuity equation are solved using an iterative method
- Finite element solutions to the Gyrokinetic-Darwin-Maxwell equations are found with multi-grid and other linear solvers



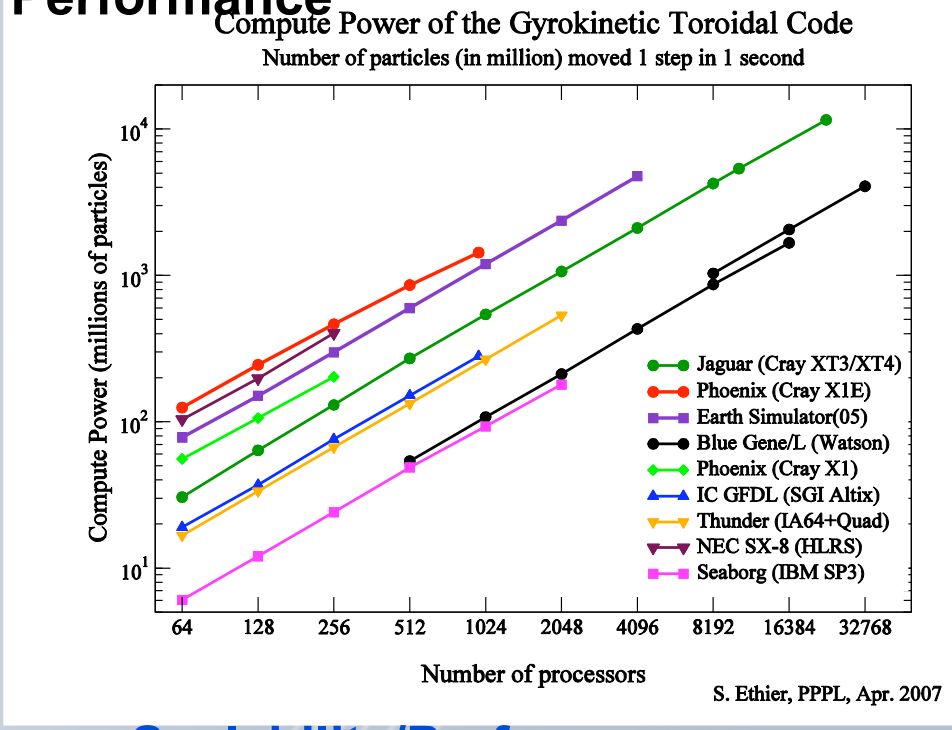
Pioneering Application: GTC

Code Readiness, Scalability, and Performance

Readiness Activities

- Physical Models
  - Implement split-weight scheme for kinetic electrons in shaped plasma component (GTC-S)
- Algorithms
  - Port and optimize GTC-S
- Scalability & performance
  - Implement radial and particle domain decomposition in GTC-S
  - Implement asynchronous I/O
  - Data flow automation
  - Joule metric benchmark studies

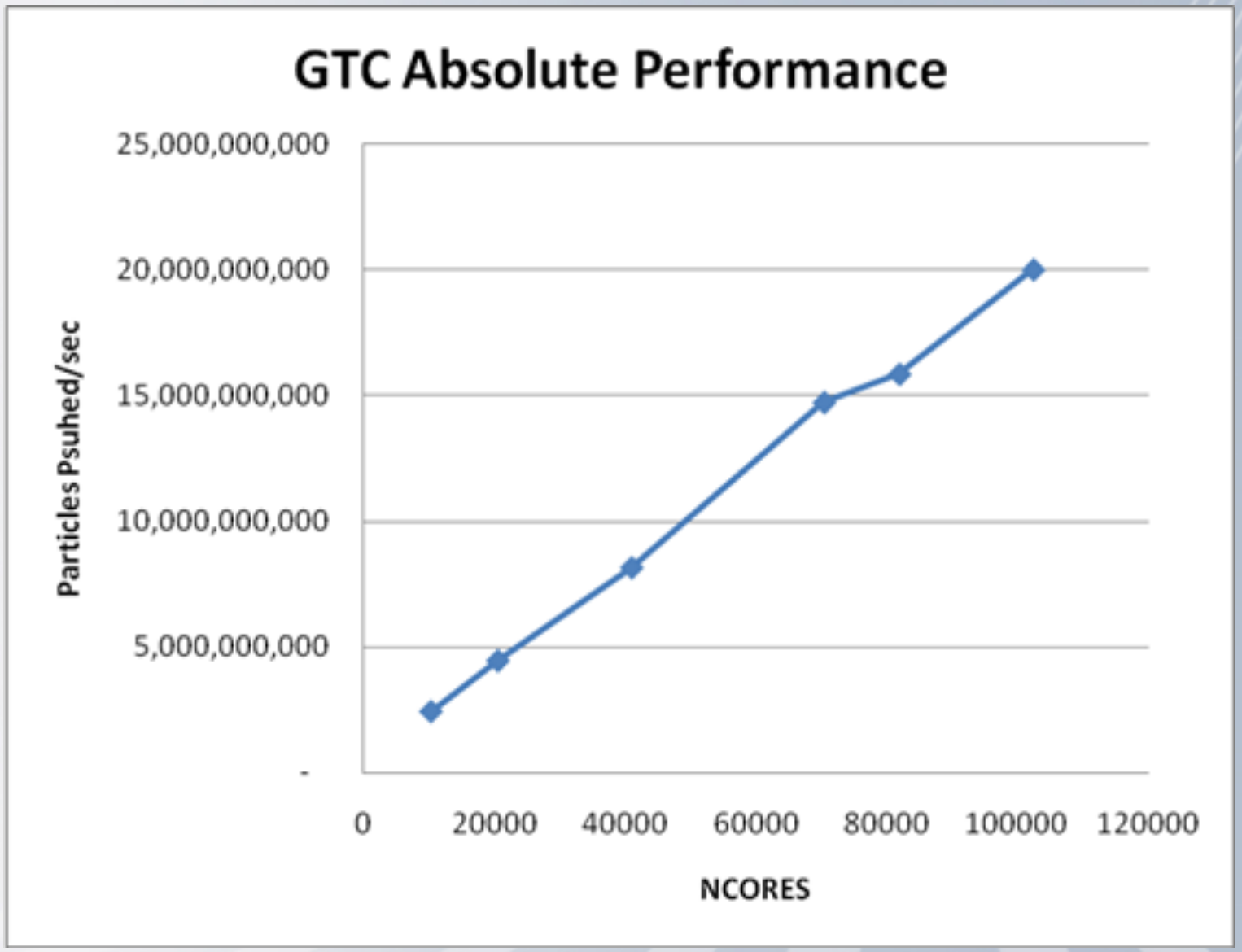
LCF liaison contributions
- **Asynchronous I/O**
- **Automated end-to-end workflow**
- **Porting/scaling new shaped plasma version**

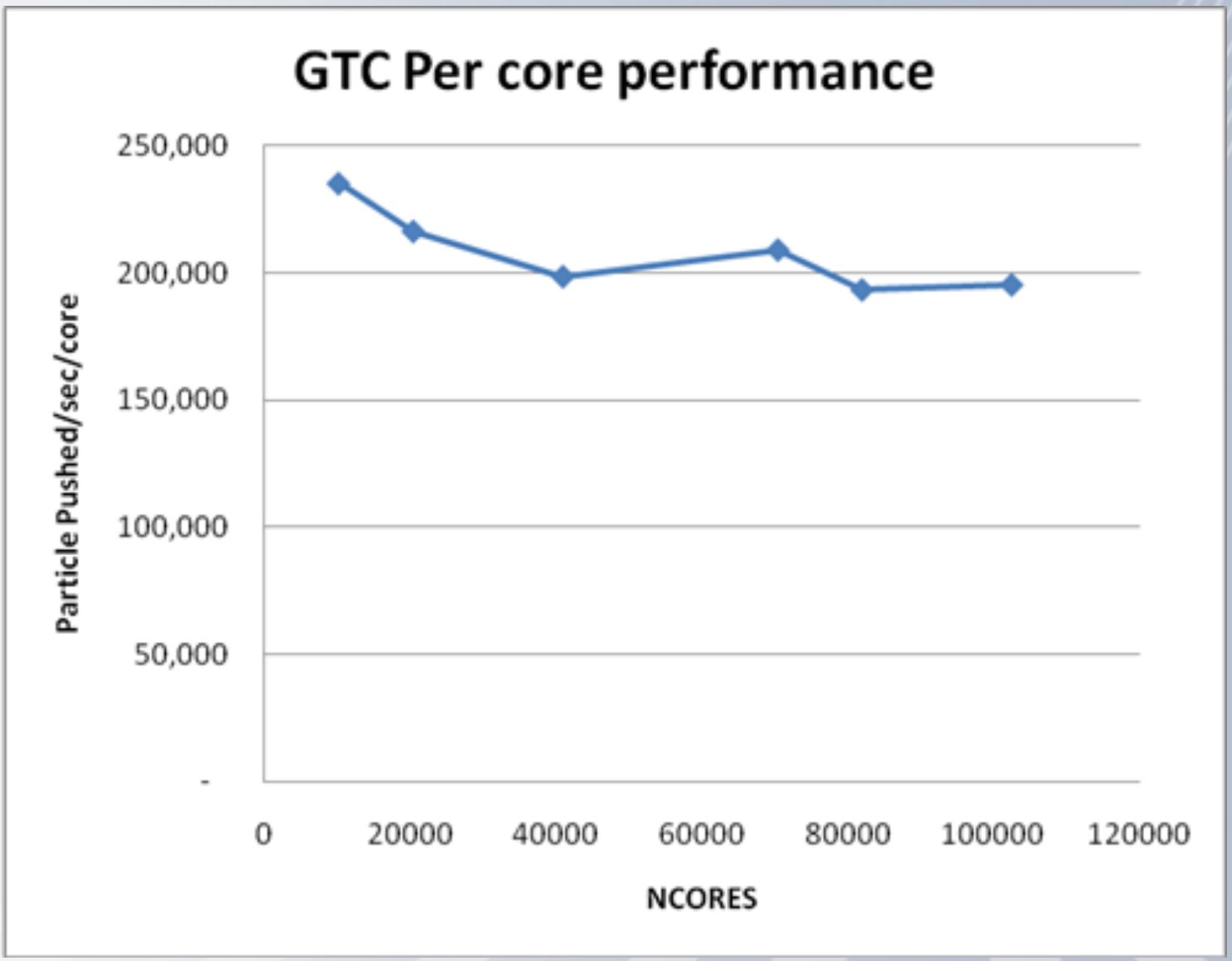


Scalability/Performance

- **Excellent full system weak scaling with ~20% of peak performance realized**
  - **Parallelized with MPI and OpenMP**
- **Initial Barcelona quad-core testbed performance promising**
  - **OpenMP threads perform well**
  - **Reduced memory B/W may not be an issue**
- **Needs to vectorize better**

Run at 142000 failed due to Code Limitations




GTC Per core performance
Particle Pushed/sec/core
250,000
200,000
150,000
100,000
50,000
-
0
20000
40000
60000
80000
100000
120000
NCORES

POCs
Mat Maltrud &
Phil Jones, LANL

Pioneering Application: POP

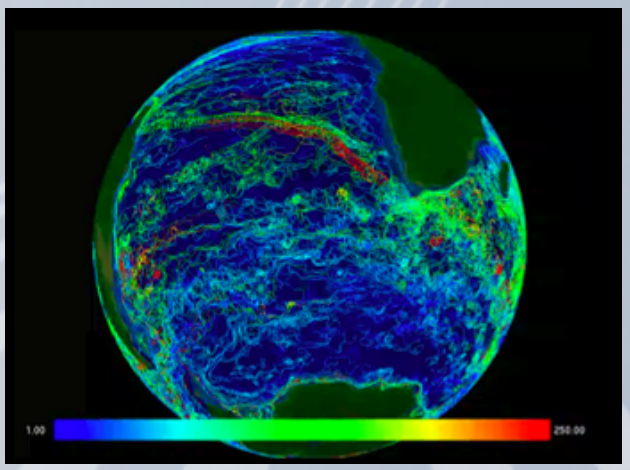
Science Goals and Impact

Science Goals

- Fundamental understanding of how the global ocean responds to the biogeochemistry feedback mechanism
  - Also facilitates model calibration in preparation for full CCSM coupling at the petascale
- Addition of biogeochemistry to the ocean model is a critical step toward prediction of the Earth system and its carbon, nitrogen, and sulphur cycles
- Simulate effects of biogeochemistry in current leading-edge eddy-resolving global ocean circulation models
  - A 20-year POP run is needed to resolve the time scales of interest
  - 0.1° resolution with tripole grid to keep coordinate singularities on land
  - Use of partial bottom cells to give more accurate bathymetry
- Sea ice model not included in current planned simulations
- 23 passive tracers will be used

Science Impact

- First-ever global eddy-resolving simulation *with* ocean biogeochemistry
  - A number of regional studies (Ross sea, Arabian Sea) have been performed but nothing global finer than 1°
- Combine the most realistic ocean simulation with a comprehensive ocean ecosystem and trace gas model
  - First attempt at a realistic simulation of ocean ecosystems
  - Include eddy pumping of nutrients and realistic simulation of fronts that are necessary for ocean ecology

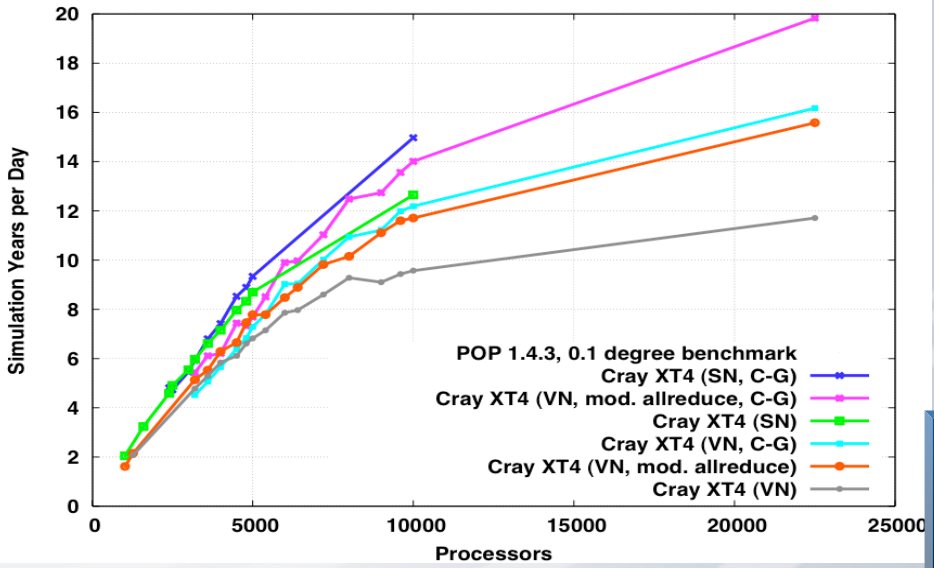


Pioneering Application: POP

Code Readiness, Scalability, and Performance

Readiness Activities

- Algorithms
  - Implement more scalable barotropic solver with improved CG preconditioner
    - Block Jacobi (additive Schwartz), with plans for multi-level enhancement
    - Trade extra flops for more iterations
- Scalability & performance
  - Tune for SSE and OpenMP parallelism
  - Implement parallel I/O and test



Scalability/Performance

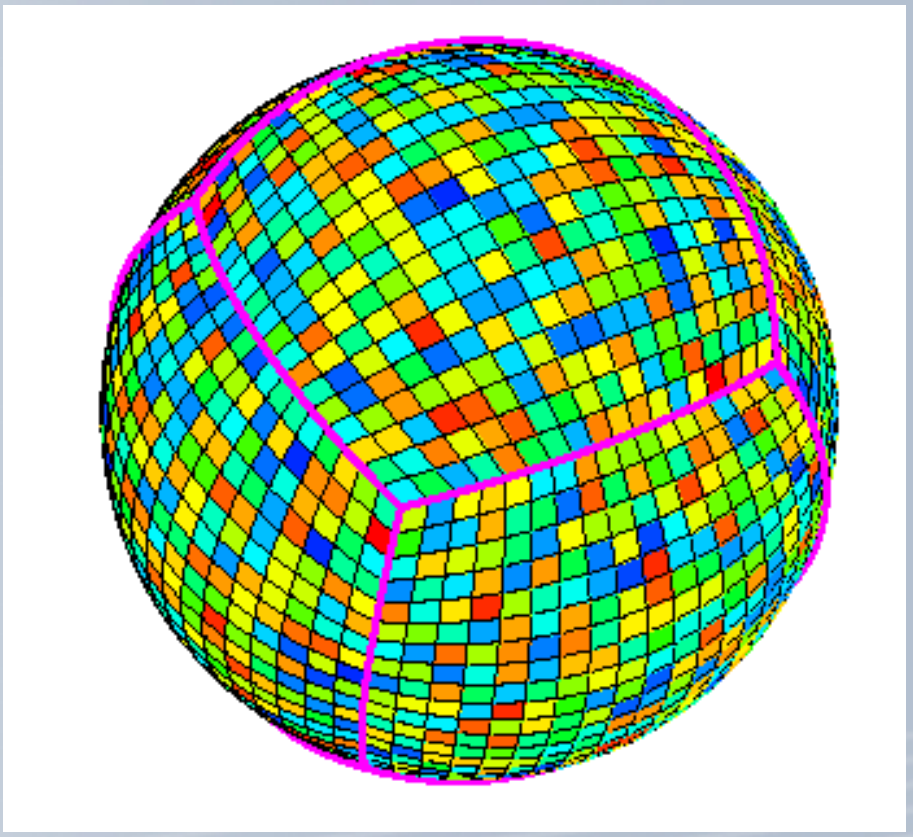
- **Ever-improving strong scaling with ~10% of peak performance**
  - **Tackle scalability-limiting barotropic solver dominated by MPI all-reduce latency with new block Jacobi preconditioner**
  - **Should benefit more from QC SSE instructions**
- **New preconditioner in barotropic solve is 1.78x faster on 15,000 cores**
  - **Full benchmark 1.38x faster**
- **Initial Barcelona quad-core testbed perf**
  - **Good vectorization**
  - **Memory B/W an issue unless high processor counts are used to ensure small subgrid size**
  - **Improved speedup needed w/ OpenMP threads**
- **Addition of biogeochemistry creates more independent work, improving scalability**
- **Issue with global gather for I/O on CNL**
  - **Currently being addressed in multiple ways**

LCF liaison contributions
- **New preconditioner for barotropic solver**
- **Contributed bug fixes to POP 2.0**
- **Represent needs at OBER/ESNET meeting**

High-Frequency Simulations of Global Seismic Wave Propagation

- A seismology challenge: model the propagation of waves near 1 hz (1 sec period), the highest frequency signals that can propagate clear across the Earth.
- These waves help reveal the 3D structure of the Earth's “enigmatic” core and can be compared to seismographic recordings.

- The Gordon Bell Team: Laura Carrington, Dimitri Komatitsch, Michael Laurenzano, Mustafa Tikir, David Michéa, Nicolas Le Goff, Allan Snavely, Jeroen Tromp



The cubed-sphere mapping of the globe represents a mesh of 6 x 182 = 1944 slices.

CRAY
THE SUPERCOMPUTER COMPANY

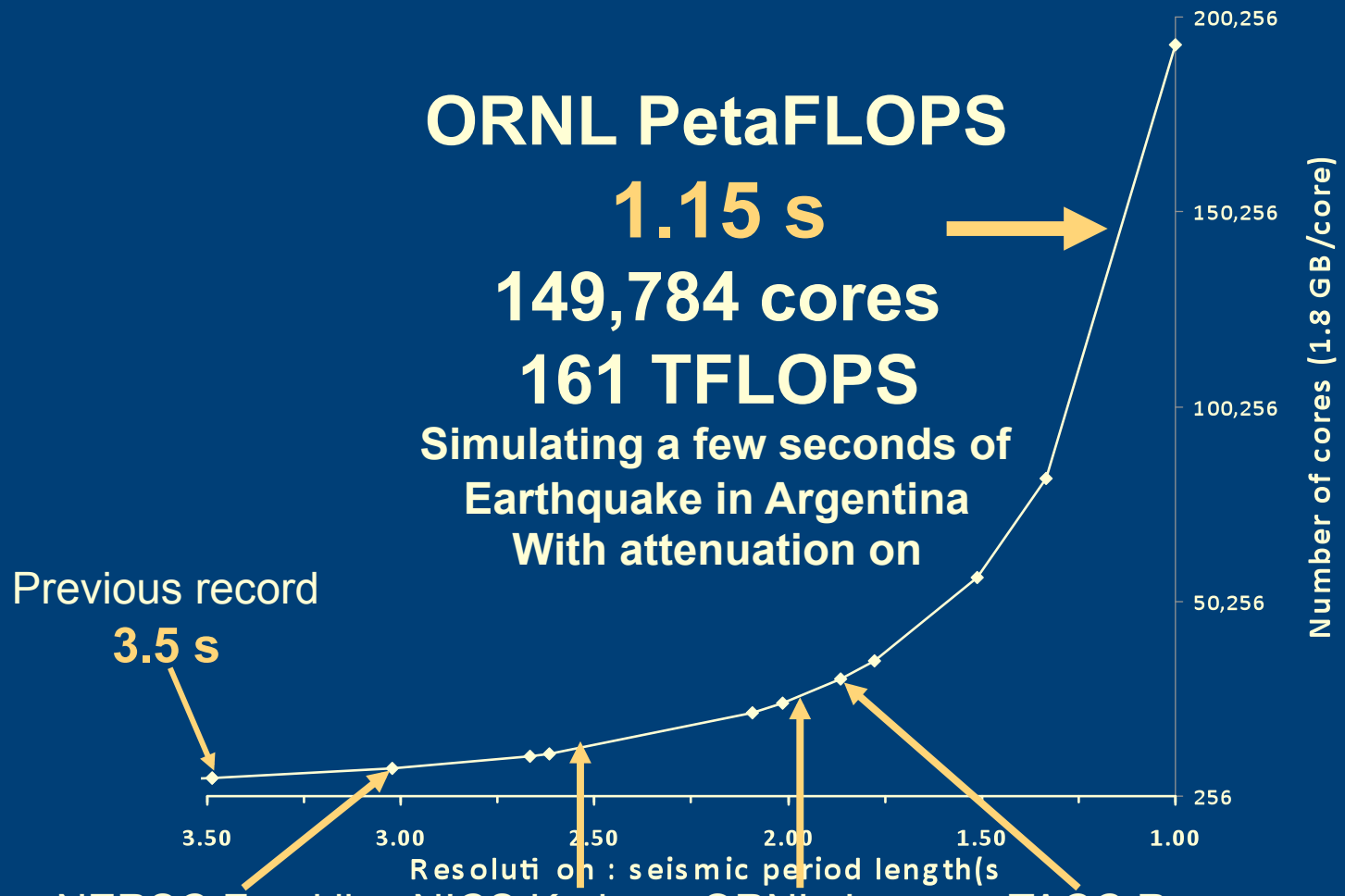
1 slide summary

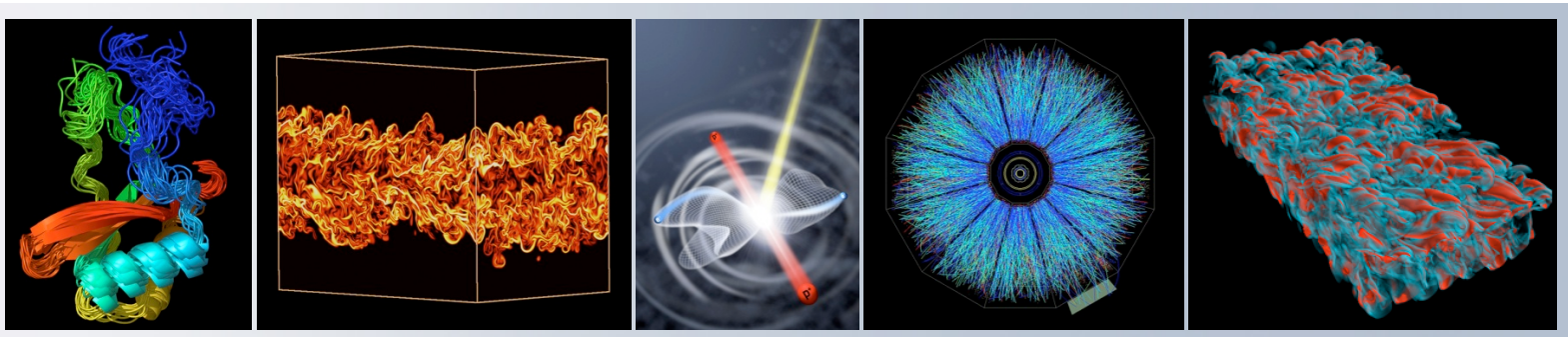
- SPECFEM3D_GLOBE is a spectral-element application enabling the simulation of global seismic wave propagation in 3D anelastic, anisotropic, rotating and self-gravitating Earth models at unprecedented resolution.
- A fundamental challenge in global seismology is to model the propagation of waves with periods between 1 and 2 seconds, the highest frequency signals that can propagate clear across the Earth.
- These waves help reveal the 3D structure of the Earth's deep interior and can be compared to seismographic recordings.
- We broke the 2 second barrier using the 32K processors of Ranger system at TACC reaching a period of 1.84 seconds with sustained 28.7 Tflops.
- We obtained similar results on the XT4 Franklin system at NERSC and the XT4 Kraken system at University of Tennessee Knoxville, while a similar run on the 28K processor Jaguar system at ORNL, which has more memory per processor, sustained 35.7 Tflops (a higher flops rate) with a 1.94 shortest period.
- This work is a finalist for the 2008 Gordon Bell Prize

Go to Gordon Bell Talk to hear about Jaguar PF results

CRAY
THE SUPERCOMPUTER COMPANY

Results - on the road to a PetaFLOPS system





Linearly Scaling Three Dimensional Fragment Method for Large Scale Electronic Structure Calculations

Lin-Wang Wang[1,2], Byounghak Lee[1], Zhengji Zhao[2], Hongzhang Shan[1,2], Juan Meza[1], David Bailey[1], Erich Strohmaier[1,2]

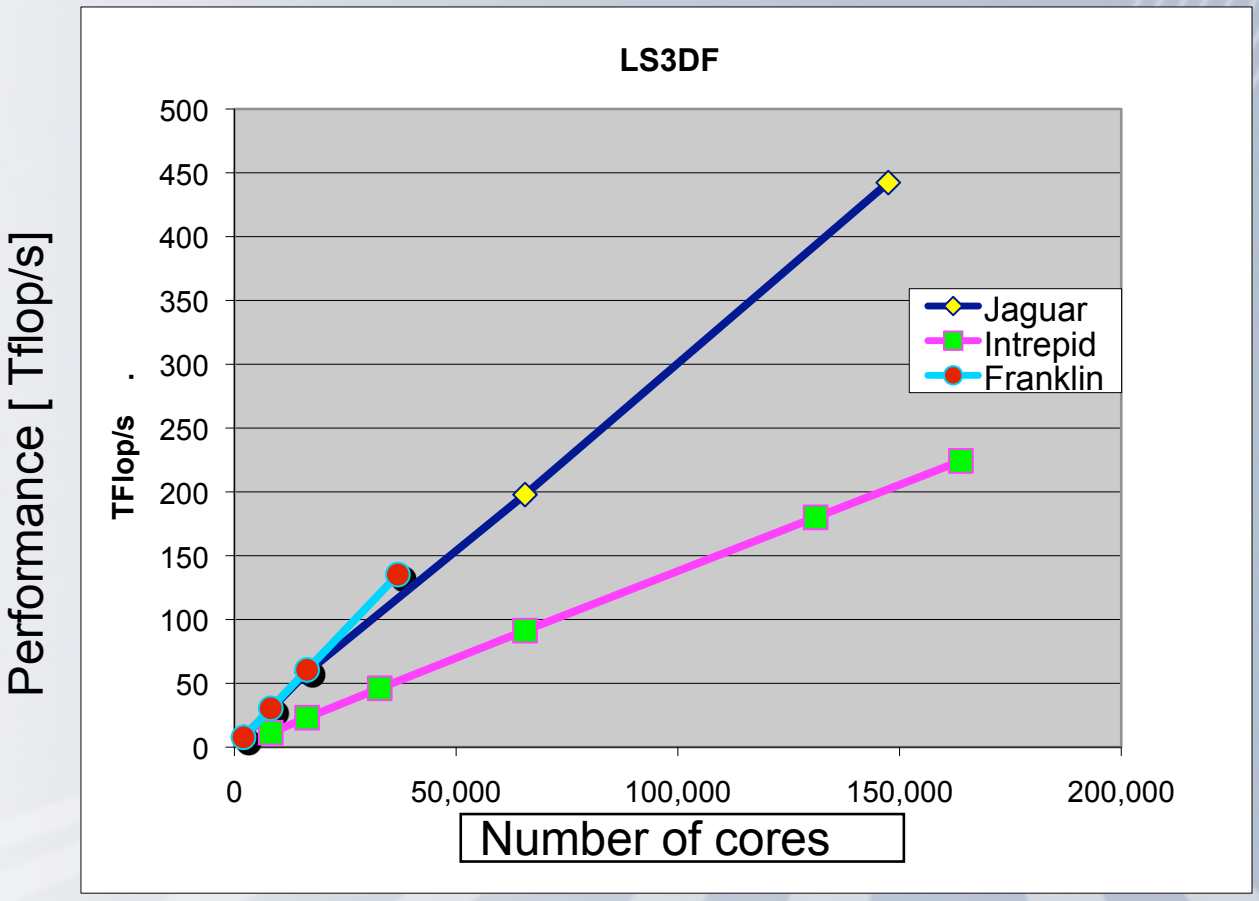
[1)]Computational Research Division
[2)]National Energy Research Scientific Computing Center (NERSC)
Lawrence Berkeley National Laboratory

US Department of Energy, Office of Science
Basic Energy Sciences and Advanced Scientific Computing Research

CRAY
THE SUPERCOMPUTER COMPANY

ZnTeO alloy weak scaling calculations

- First large scale run on Franklin at NERSC: 135 Tflops, 40% efficiency
- Subsequent runs on Intrepid at ALCF: 224 Tflops, 40% efficiency
- Final runs on Jaguar XT5 at NCCS: 442 Tflops, 33% efficiency



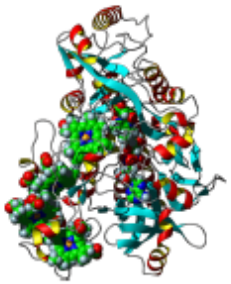
Note: Ecut = 60Ryd with *d* states, up to 36864 atoms

Transition to Operations

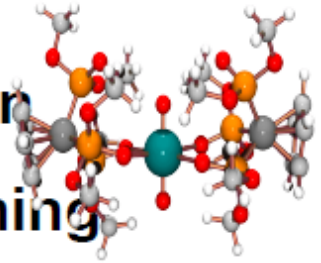
A New Period After Acceptance, Before General Availability

- ORNL LCF systems enter a Transition to Operations (T2O) period
  - Upon passing Acceptance in that system's Acceptance Test Plan
  - A short period pre-negotiated with DOE ASCR Program Management
- The T2O has three principal goals
  - Achieve at-scale "science on day one" with early access pioneering apps
  - Address any outstanding system problems found during acceptance
  - Subject system to a real production workload, thereby increasing stability
- The T2O period is a "limited availability" period
  - Those users associated with the pioneering applications have higher job and reservation priority
- The actual T2O phase for a given LCF system
  - Lasts for a period that depends upon pre-defined completion criteria
    - The criteria for completion is system dependent and negotiated in advance
- System enters General Availability after the T2O period
  - All INCITE users allowed on system at this time
- T2O plans are documented in advance for each LCF system
  - T2O Execution Plan for the PFLOP system is available

NWChem



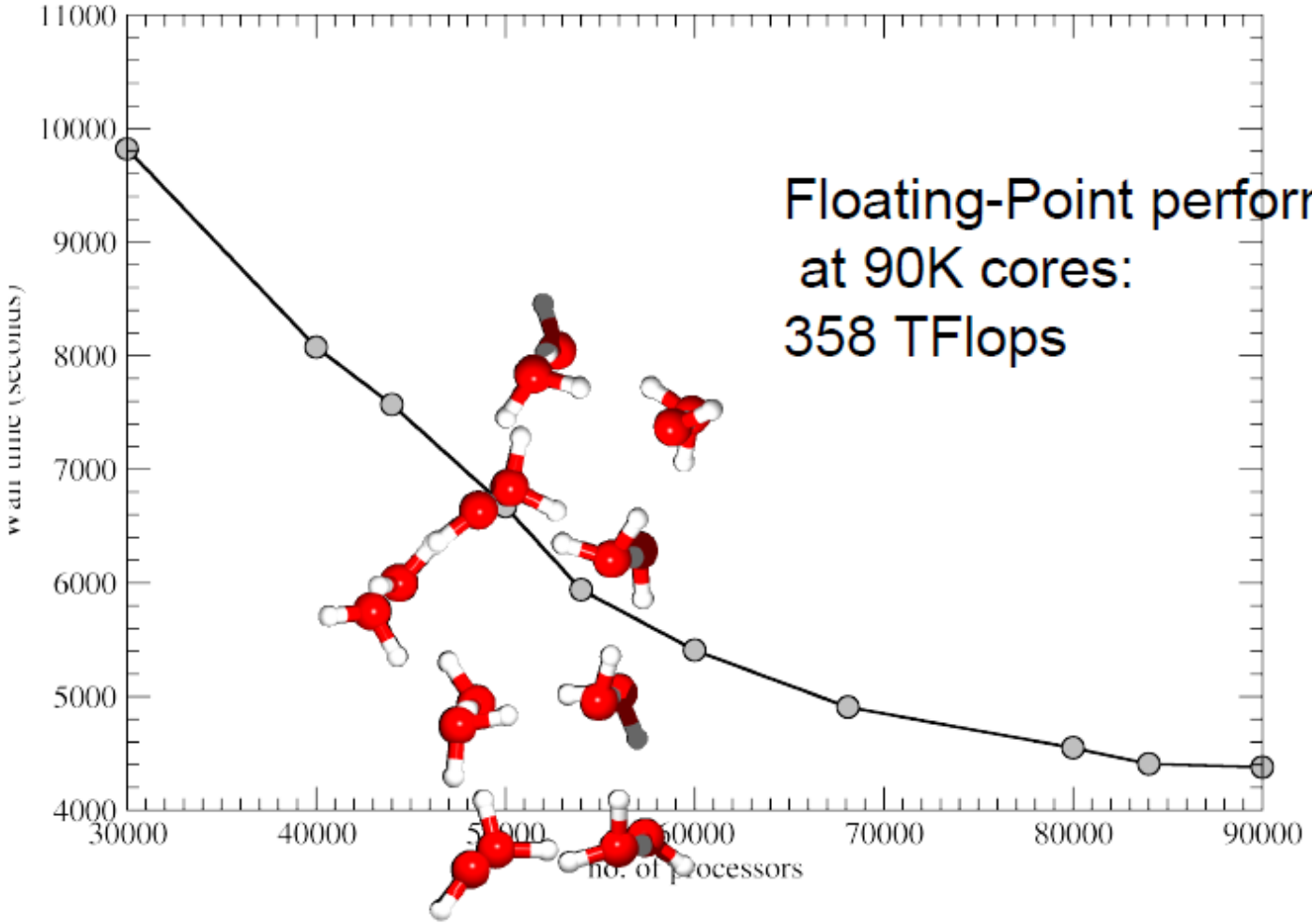
- Premier DOE computational chemistry software developed at PNNL
- Provides science to solution
- Provide major modeling and simulation capability for molecular science
  - Broad range of molecules, including catalysists, biomolecules, and heavy elements
  - Solid state capabilities
- Performance characteristics – designed for MPP
  - Single node performance comparable to best serial codes
  - Runs on a wide range of computers
- Uses Global Arrays/ARMCI for parallelization
- ORNL contribution: Cray XT porting and tuning



New ARMCI port for Cray XT

- Part of collaborative effort between ORNL and PNNL. This ORNL contributed port will be integrated in next GA/NWChem releases distributed by PNNL
- Uses the Portals library for inter-node communication
- SMP aware: uses SysV shared memory for intra-node communication
- Server-layer for calls that do not map directly onto the network

CCSD(T) run on Cray XT5 : 18 water

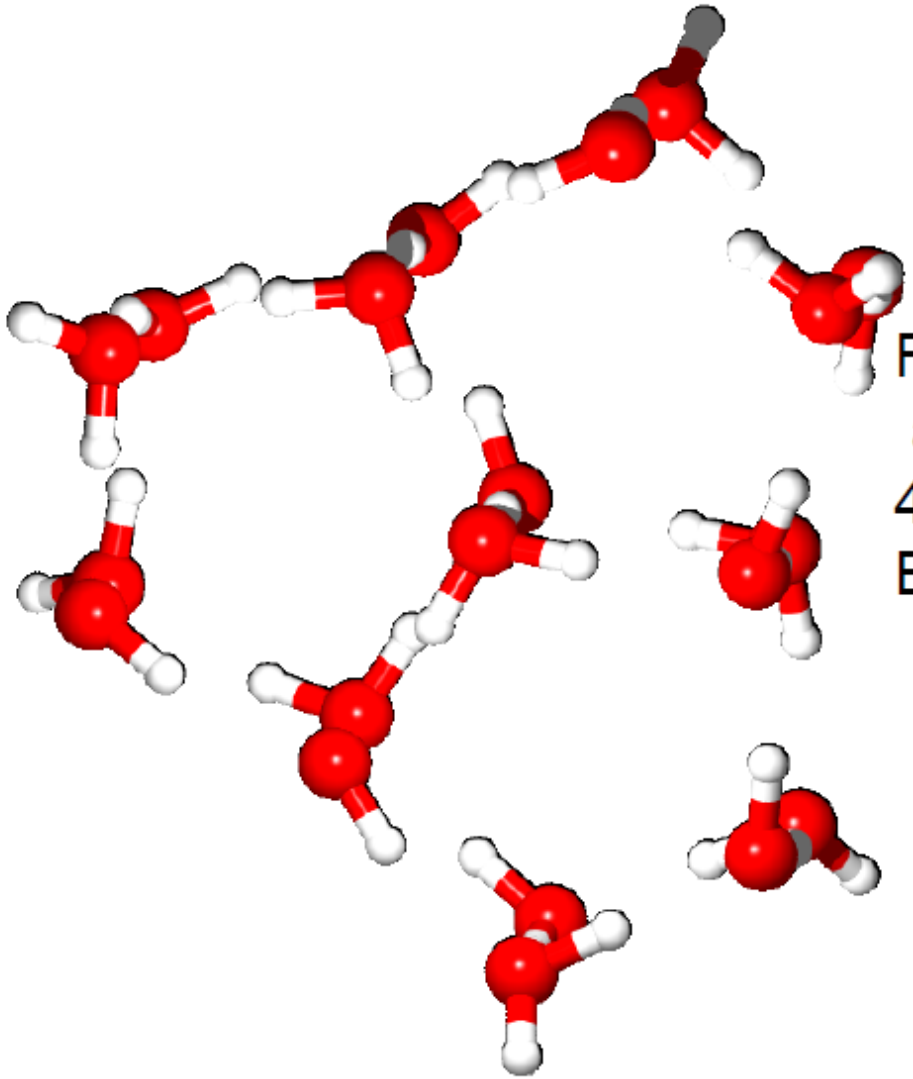


Floating-Point performance
at 90K cores:
358 TFlops

$(H_2O)_{18}$

54 atoms
918 basis functions
Cc-pvtz(-f) basis

CCSD(T) run on Cray XT5 : 20 water



Floating-Point performance
at 92K cores:
475 TFlops
Efficiency > 50%

$(H_2O)_{20}$

60 atoms
1020 basis functions
Cc-pvtz(-f) basis

CP2K

- -
- REFERENCES -
- -

CP2K version 2.0.1 (Development Version), the CP2K developers group (2009).
CP2K is freely available from http://cp2k.berlios.de/ .

Guidon, M; Schiffmann, F; Hutter, J; VandeVondele, J.
JOURNAL OF CHEMICAL PHYSICS, 128 (21), 214104 (2008).
Ab initio molecular dynamics using hybrid density functionals.
http://dx.doi.org/10.1063/1.2931945

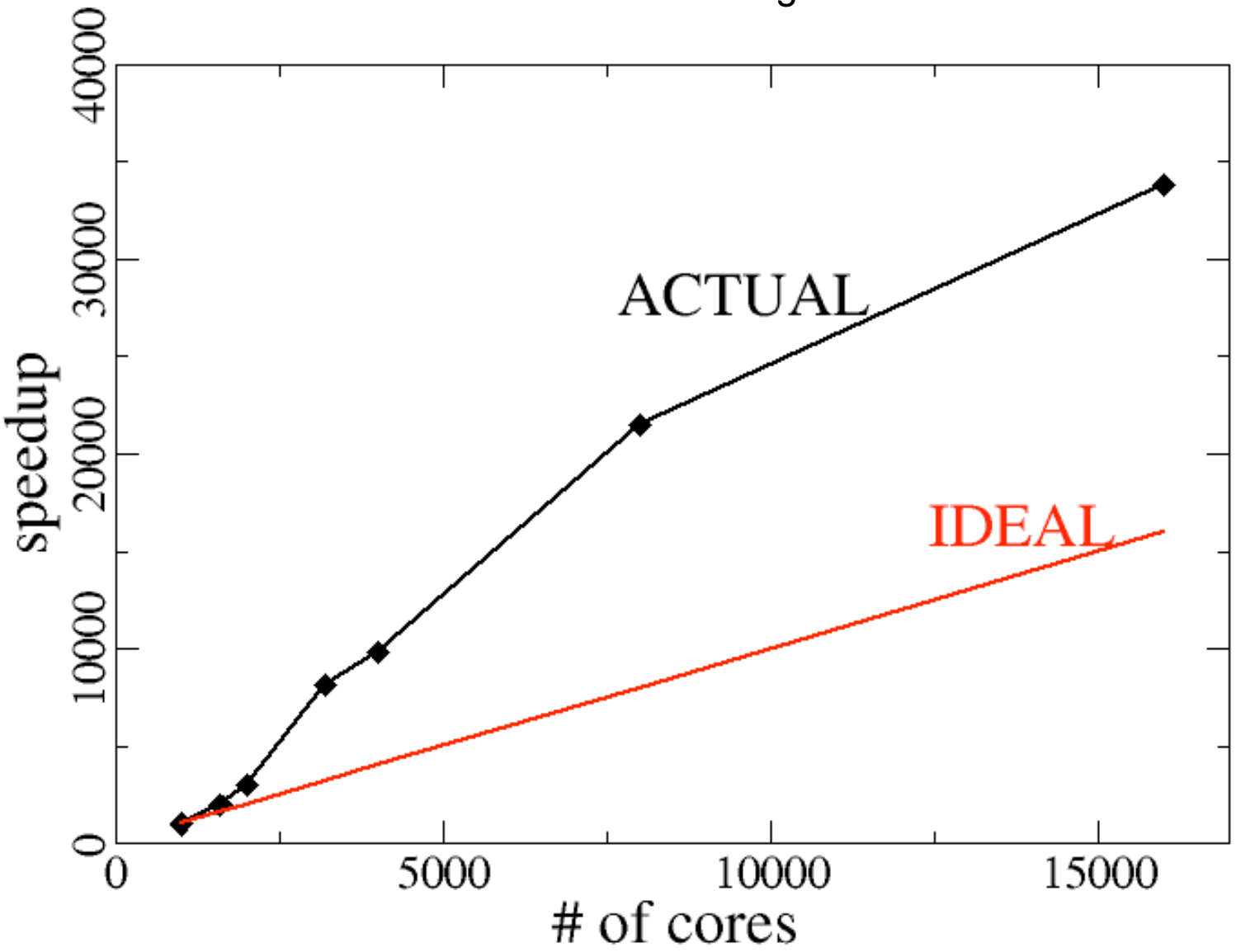
Frigo, M; Johnson, SG.
PROCEEDINGS OF THE IEEE, 93 (2), 216-231 (2005).
The design and implementation of FFTW3.
http://dx.doi.org/10.1109/JPROC.2004.840301

Kolafa, J.
JOURNAL OF COMPUTATIONAL CHEMISTRY, 25 (3), 335-342 (2004).
Time-reversible always stable predictor-corrector method for molecular dynamics of polarizable molecules.
http://dx.doi.org/10.1002/jcc.10385

VandeVondele, J; Hutter, J.
JOURNAL OF CHEMICAL PHYSICS, 118 (10), 4365-4369 (2003).
An efficient orbital transformation method for electronic structure
calculations.
http://dx.doi.org/10.1063/1.1543154

Lippert, G; Hutter, J; Parrinello, M.
THEORETICAL CHEMISTRY ACCOUNTS, 103 (2), 124-140 (1999).
The Gaussian and augmented-plane-wave density functional method for ab
initio molecular dynamics simulations.

Recent CP2K Timings



Turbulence Studies on the Cray XT

P. K. Yeung[1], D.A. Donzis[2], D. Pekurovsky[3]

[1] Georgia Tech; E-mail: pk.yeung@ae.gatech.edu
[2] Univ. of Maryland; [3] San Diego Supercomputer Center

Supported by NSF (CBET-Fluid Dynamics, OCI-PetaApps)

NICS: XT4, XT5 (TeraGrid, early user)

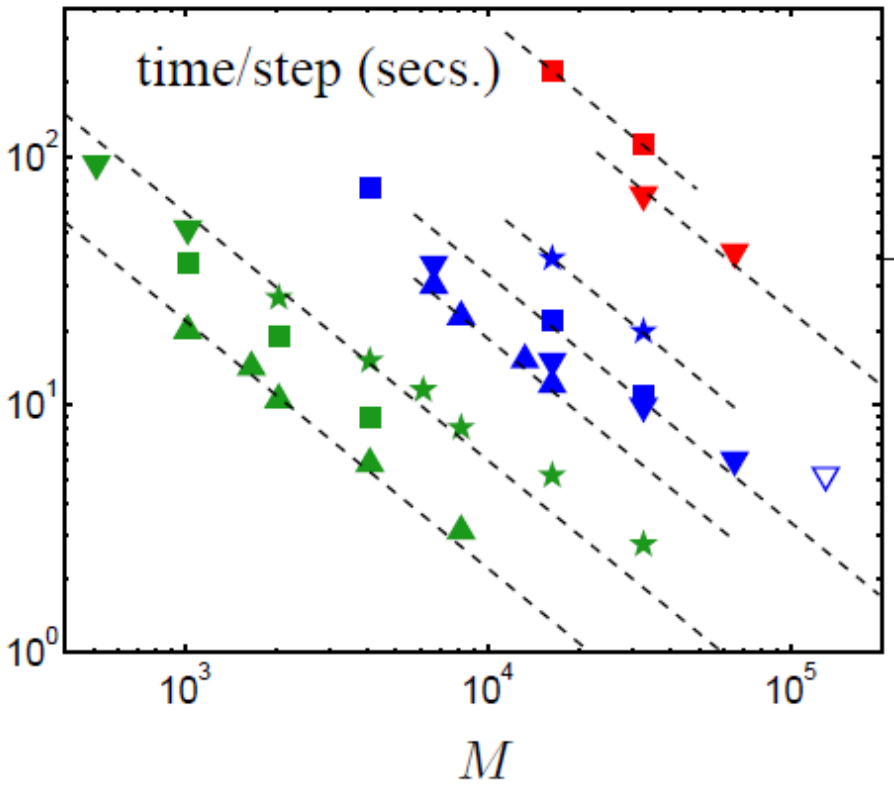
NCCS: XT5 (Petascale Early Science Project)

Cray Technical Workshop
Charleston, SC, Feb. 24-25, 2009

Thanks to the helpers:
— K. Wong (NICS), D. Kothe (NCCS); J. Larkin, J. Levesque (Cray)

Strong Scaling (Eulerian DNS only)

- For a given problem size N^3: $(t/s) \propto M^{-1}$ if perfect



$N^3 = 2048^3, 4096^3, 8192^3$

Squares: Ranger (TACC)
Stars: BlueGene L (SDSC/IBM)
Up Triangles: Cray XT4
Down Triangles: Cray XT5
($\leq$ 64k at NICS, 128k at NCCS)

- Data for best processor-grid $M_1 \times M_2$ being shown
- Continued high scalability on XT5 proven difficult (so far)
- Actual performance shows variability (due to other jobs)

WHAT ABOUT I/O

Chimera Benchmark Results

- Why ADIOS is better than pHDF5?

ADIOS_MPI_IO vs. pHDF5 w/ MPI Indep. IO driver

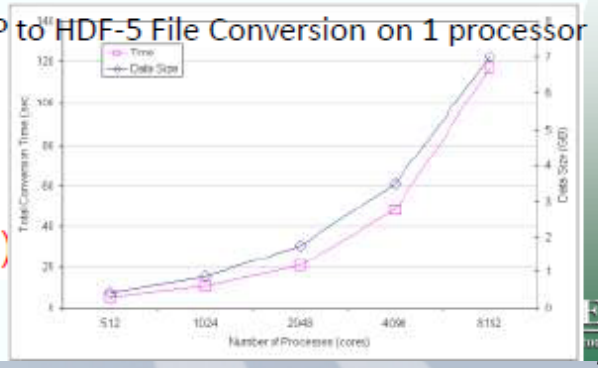
ADIOS_MPI_IO		
Function	# of calls	Time
write	2560	2218.28
MPI_File_open	2560	95.80
MPI_Recv	2555	24.68
buffer_write	6136320	10.29
fopen	512	9.86
bp_calsize_stringtag	3179520	4.44
other	--	~40

pHDF5		
Function	# of calls	Time
write	144065	33109.67
MPI_Bcast(sync)	314800	12259.30
MPI_File_open	2560	325.17
MPI_File_set_size	2560	23.76
MPI_Comm_dup	5120	16.34
H5P,H5D,etc	--	8.71
other	--	~20

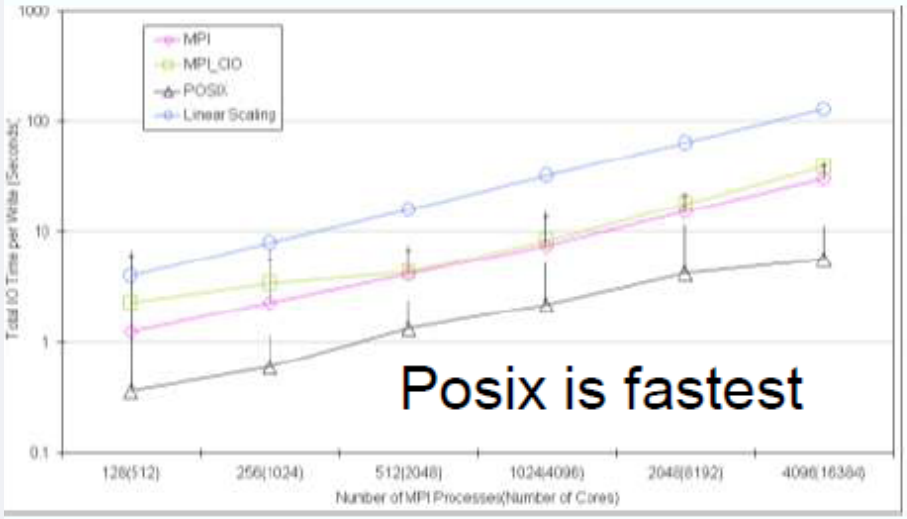
Use 512 cores, 5 restart dumps.

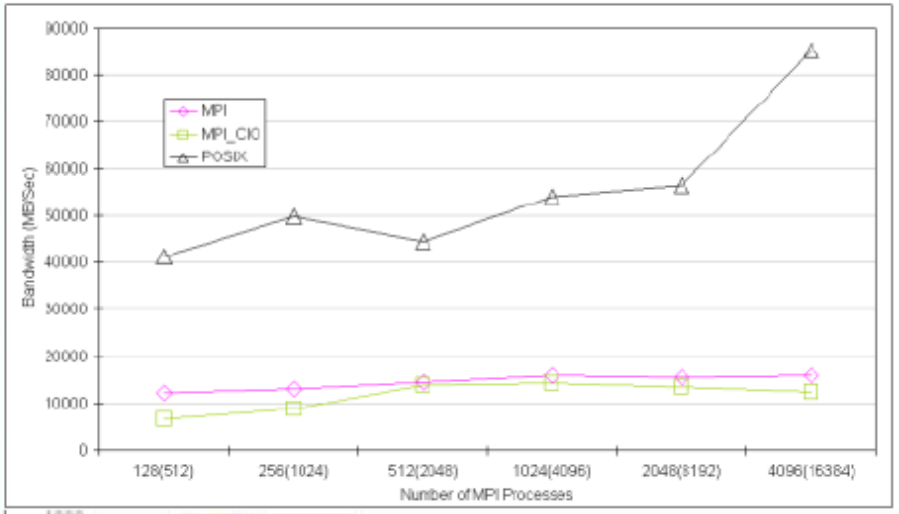
Conversion time on 1 processor for the 2048 core job =
3.6s (read) + 5.6s (write) + 6.9 (other) = 18.8 s

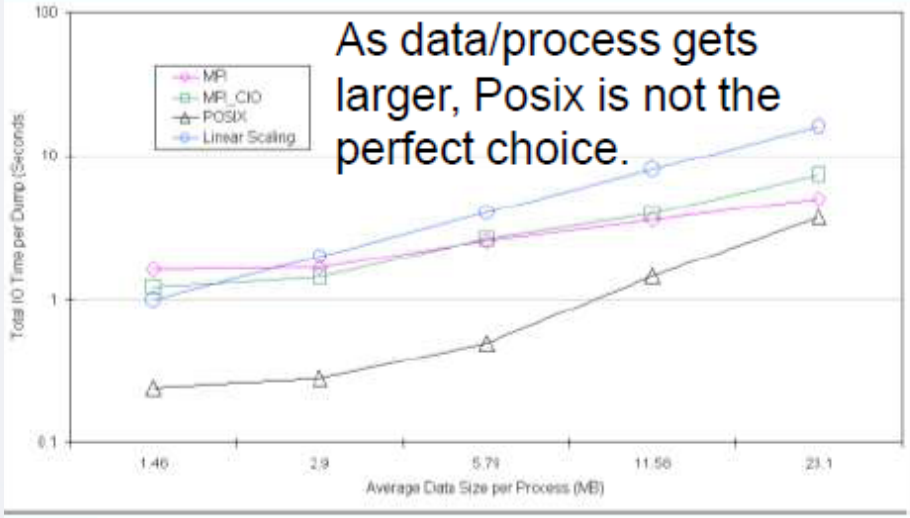
Number above are sum among all PEs (parallelism not shown)

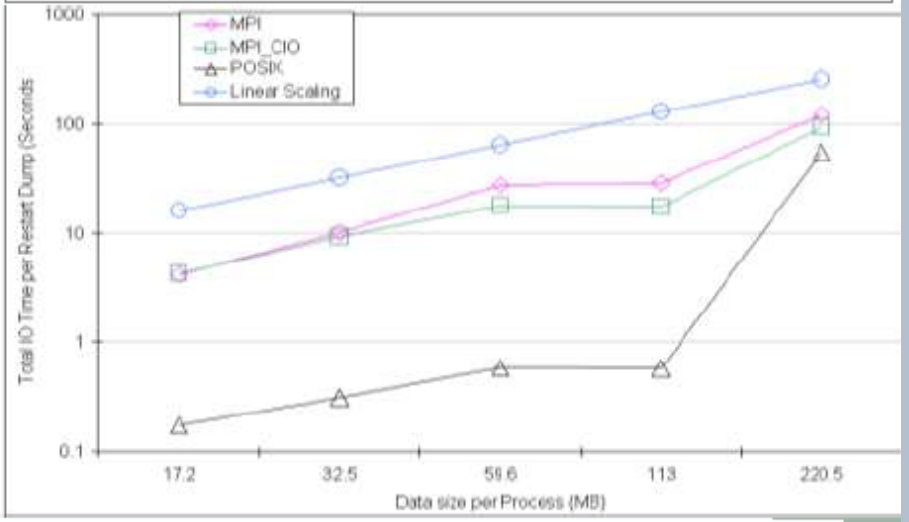


GTC Performance





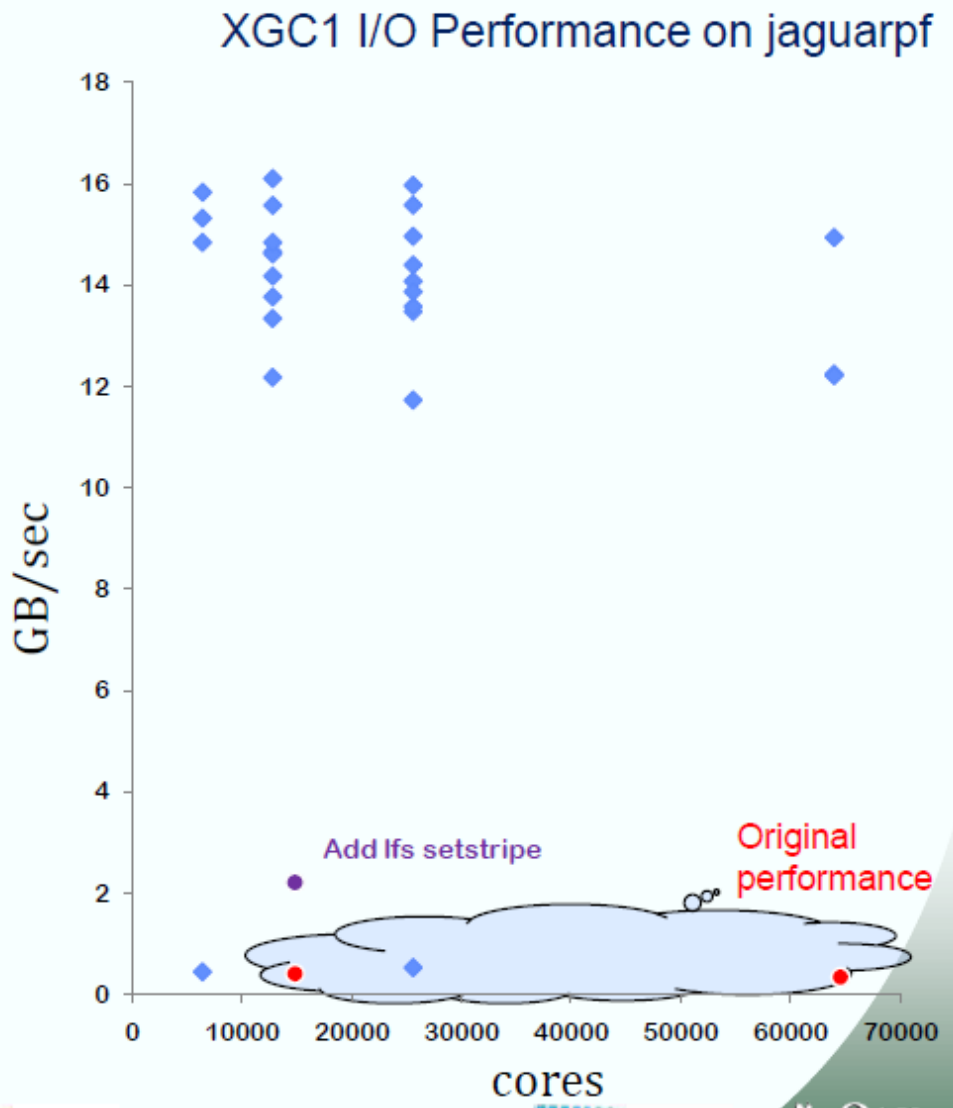




Best Performance was 32 shared files per restart write, 28,800 cores (4 openmp threads), 38GB/file on jaguar. Time to write = 49 seconds. (1.2 TB)

XGC-1 I/O performance on Jaguarpf

- Original output with 1 shared file with ADIOS-MPI I/O, no stripe alignment.
- Add lustre, lfs setstripe count to 160.
- Make nphi (16) shared files, add setstripe in ADIOS.



Summary of Applications

Application	Core Count	MPI	OpenMP	Scaling
DCA++	150000	Yes	No	Weak
LSMS	150000	Yes	No	Weak
SPECFem3D	150000	Yes	No	Weak
WRF	150000	Yes	No	Strong
S3D	144000	Yes	No	Weak
GTC	120000	Yes	Yes	Weak
LS3DF	150000	Yes	No	Weak
NWCHEM	92000	Global Arrays	No	Strong
CP2K	32000	Yes	Yes	Strong

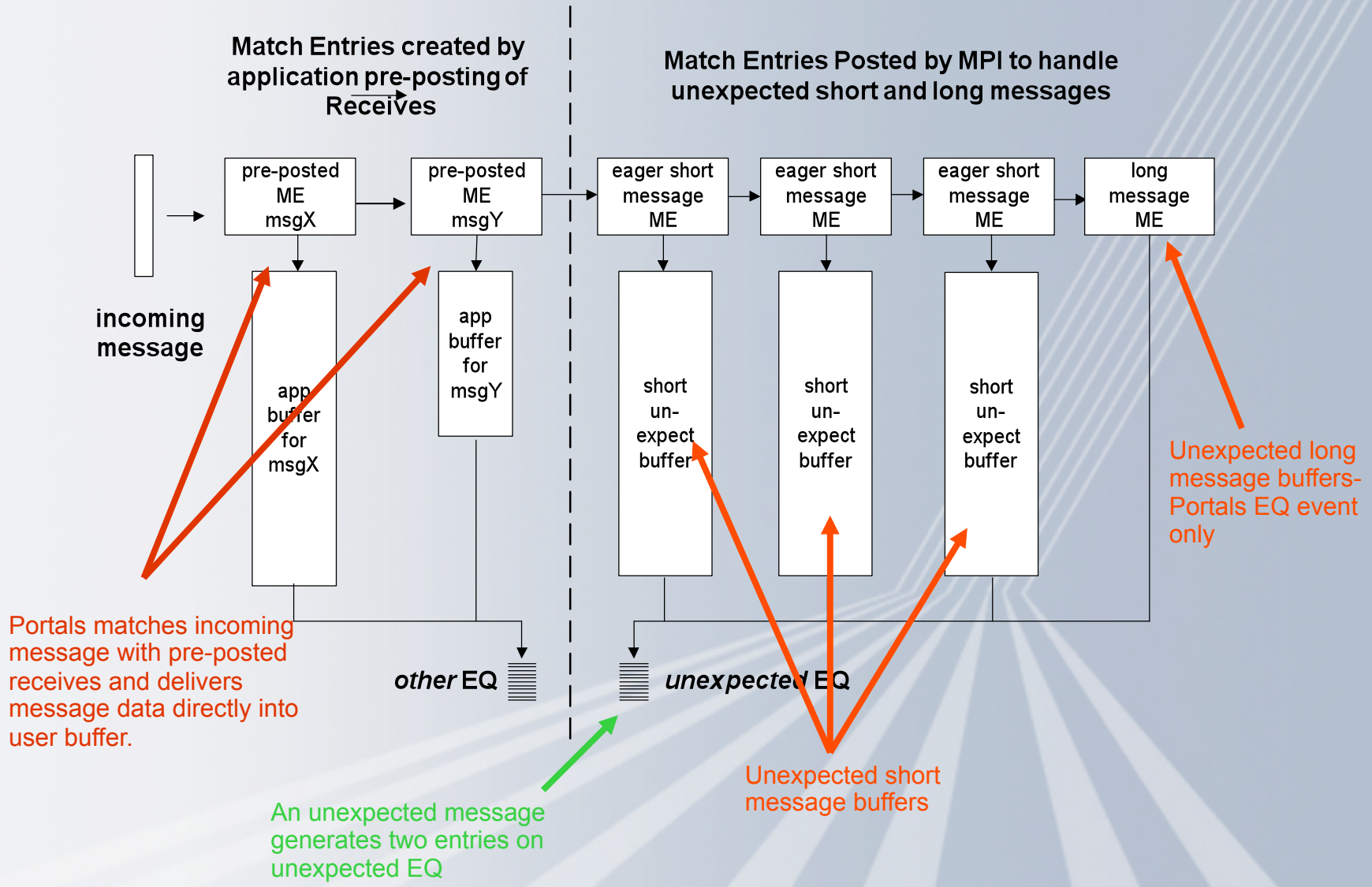
XT MPI Implementation

Two protocols are used for handling basic MPI-1 send/recv style messaging:

- Eager protocol for short messages
- Two protocols for long messages

But first we will talk about the receive side, since that is where Portal's important features with respect to MPI are most evident…

XT MPI – Receive Side



XT MPI Short Message Protocol-Sending side

- If message to send has no more than 128000 bytes of data, the short message, eager protocol is used by sender.
- Sender assumes the receiver has space for the message in an MPI-internal receive buffer and 2 slots in the *unexpected Event Queue.*
- For very short messages (1024 bytes or shorter), sender copies data to internal buffer before sending *PtlPut* request to Portals.

XT MPI-2 RMA

- XT MPI supports all RMA operations
- Based on MPICH2 CH3 device RMA
  - Layered on top of internal send/recv protocol
- Designed for functionality, not performance.
- Little opportunity for overlapping of communication with computation when using MPI-2 RMA on XT.
- Almost all communication occurs at end of exposure epochs or in *MPI_Win_free.*

What does this mean? (1)

If you see this error message:

```
internal ABORT - process 0: Other MPI error, error stack:
MPIDI_PortalsU_Request_PUPE(317): exhausted unexpected receive queue
buffering increase via env. var. MPICH_UNEX_BUFFER_SIZE
```

It means:

The application is sending too many short, unexpected messages to a particular receiver.

Try doing this to work around the problem:

Increase the amount of memory for MPI buffering using the MPICH_UNEX_BUFFER_SIZE variable(default is 60 MB) and/or decrease the short message threshold using the MPICH_MAX_SHORT_MSG_SIZE (default is 128000 bytes) variable. May want to set MPICH_DBMASK to 0x200 to get a traceback/coredump to learn where in application this problem is occurring.

What does this mean? (2)

If you see this error message:

```
Assertion failed in file /notbackedup/users/rsrel/
rs64.REL_1_4_06.060419.Wed/pe/computelibs/mpich2/src/mpid/portals32/
src/portals_init.c at line 193: MPIDI_Portals_unex_block_size >
MPIDI_Portals_short_size
```

It means:

The appearance of this assertion means that the size of the unexpected buffer space is too small to contain even 1 unexpected short message.

Try doing this to work around the problem:

User needs to check their MPICH environment settings to make sure there are no conflicts between the setting of the MPICH_UNEX_BUFFER_SIZE variable and the setting for MPICH_MAX_SHORT_MSG_SIZE. Note setting MPICH_UNEX_BUFFER_SIZE too large (> 2 GB) may confuse MPICH and also lead to this message.

What does this mean? (3)

If you see this error message:

```
[0] MPIDI_PortalsU_Request_FDU_or_AEP: dropped event on unexpected
receive queue, increase

[0] queue size by setting the environment variable MPICH_PTL_UNEX_EVENTS
```

It means:

You have exhausted the event queue entries associated with the unexpected queue. The default size is 20480.

Try doing this to work around the problem:

You can increase the size of this queue by setting the environment variable MPICH_PTL_UNEX_EVENTS to some value higher than 20480.

What does this mean? (4)

If you see this error message:

```
[0] MPIDI_Portals_Progress: dropped event on "other" queue, increase
[0] queue size by setting the environment variable MPICH_PTL_OTHER_EVENTS
aborting job: Dropped Portals event
```

It means:

You have exhausted the event queue entries associated with the "other" queue. This can happen if the application is posting many non-blocking sends, or a large number of pre-posted receives are being posted, or many MPI-2 RMA operations are posted in a single epoch. The default size of the other EQ is 2048.

Try doing this to work around the problem:

You can increase the size of this queue by setting the environment variable MPICH_PTL_OTHER_EVENTS to some value higher than the 2048 default.

What does this mean? (5)

If you see this error message:

```
0:(/notbackedup/users/rsrel/rs64.REL_1_3_12.051214.Wed/pe/
computelibs/mpich2/src/mpid/portals32/src/portals_progress.c:642)
PtlEQAlloc failed : PTL_NO_SPACE
```

It means:

You have requested so much EQ space for MPI (and possibly SHMEM if using both in same application) that there are not sufficient Portals resources to satisfy the request.

Try doing this to work around the problem:

You can decrease the size of the event queues by setting the environment variable MPICH_PTL_UNEXPECTED_EVENTS and

MPICH_PTL_OTHER_EVENTS to smaller values.

What does this mean? (6)

If you see this error message:

```
aborting job: Fatal error in MPI_Init: Other MPI error, error
stack: MPIR_Init_thread(195): Initialization failed
MPID_Init(170): failure during portals initialization
MPIDI_Portals_Init(321): progress_init failed
MPIDI_PortalsI_Progress_init(653): Out of memory
```

It means:

There is not enough memory on the nodes for the program plus MPI buffers to fit.

Try doing this to work around the problem:

You can decrease the amount of memory that MPI is using for buffers by using MPICH_UNEX_BUFFER_SIZE environment variable.

Additional Tidbits

- Tripolar grid in POP
  - Changed CNL default from SMP to distributed placement
    - Factor of two in runtime
- DCA++
  - MPI_ALLTOALL across all n*2048 clusters changed to MPI-ALLTOALL across n with broadcast to 2048 members of cluster
    - Enabled scaling to 149000
- Several misc codes
  - Combined 8-10 MPI_ALLTOALLS to one MPI_ALLTOALL with an array of values

Next Generation XT

- 12 Core Nodes
  - Istanbul with improved memory controller
- Gemini interconnect
  - Enable PGAS implementations
  - 100 times the MPI messages/second
  - Very low latency